AI for Self-Assessment: Facial Emotion-Tracking in Video Mock Interviews

Vanessa Sanchez
University of Texas at Austin
vanessa.sanchez@utexas.edu

Silvia DalBen Furtado
University of Texas at Austin
silviadalben@utexas.edu

Dhanny Indrakusuma
University of Texas at Austin
dhannywi@utexas.edu

Kyle Soares
University of Texas at Austin
kyle.soares@utexas.edu

ABSTRACT

AI's challenges with transparency and explainability have become engrained through all stages of the hiring process in the last decade. In this study, we designed a mock interview experiment to quantify the impact of AI-driven facial emotion recognition. Are those AI systems able to consistently and accurately measure emotions and objectively deduce behaviors from emotion-tracking data? We conducted nine remote mock interviews and analyzed the answers using an open source Facial Expression Recognition (FER) model on Python used for sentiment analysis of images and videos. We curated individualized analytics to understand the impact of AI emotion-tracking on video interviews and how such tools can be used for effective mock video interview preparation. While facial recognition adds complexity and stress in interview settings, emotion-tracking outputs can be used for increased self awareness in behavioral interviews. We hope to empower people interviewed with AI and encourage transparency and helpful feedback loops from AI interview-prep companies.

Keywords
Artificial Intelligence ● *Facial Recognition* ● *Emotion-Tracking* ● *Video Mock Interview* ● *Hiring*

I. Introduction

Background and relevance

AI's challenges with transparency and explainability have become engrained through all stages of the hiring process in the last decade. These hiring shifts have significantly and disproportionately impacted cultural minorities and disadvantaged communities. Simultaneously, numerous AI-driven resources have been developed to aid job seekers navigate these black-box hiring obstacles.

In a race to select the best talents, companies, HR professionals and recruiters have increasingly adopted AI hiring technologies from screening resumes to conducting virtual interviews, using systems such as *HireVue*. This controversial tool recorded the facial expressions of interviewees in virtual interviews and shared with employers calculated correlations of behavioral traits and characteristics based on those expressions. *HireVue* discontinued its "facial analysis component from its screening assessments" in 2020, but the prediction of behavioral traits and characteristics still exists within the platform and it is now handled by separate natural language processing algorithms in the software.

In this study, we designed a mock interview experiment to study the factors of using AI-driven facial emotion recognition in practice interviews. We curated individualized analytics to understand the impact of AI emotion-tracking on individual video interviews and how such tools can be used for effective mock video interview preparation.

Objective

The purpose of our study is to quantify the impact of AI emotion-tracking used in video interviews and to visualize helpful emotion-tracking reports. We want to reflect upon how AI could be used for good in mock interview settings and to document the limitations of using emotion tracking in a video interview, as the use of AI has become a commonplace throughout the hiring process.

Following are the 4 research questions we aim to answer:

RQ1: Are those AI systems able to consistently and accurately measure emotions and objectively deduce behaviors from emotion-tracking data?

RQ2: Do they accurately analyze users with darker skin tones?

RQ3: How does the knowledge of AI emotion-tracking in mock video interviews influence user performance?

RQ4: Could AI emotion recognition be used as a diagnostic tool to increase facial emotional expression awareness in interview settings?

What we did

We conducted nine remote mock interviews recorded on Zoom with participants recruited through a screener survey. Some of our participants were actively searching for jobs while others were not. They were all aware that this was an experiment focused on analyzing their facial expressions during their responses. We asked them to pretend they were

participating in an interview for their ideal job. We followed a script with three behavioral questions designed to elicit neutrality (baseline), confidence and stress, followed by three post-interview questions.

Each mock interview video was analyzed by an open source Facial Expression Recognition (FER) model on Python used for sentiment analysis of images and videos. It generated a .csv file with seven emotions - neutral, happy, surprise, sad, anger, fear and disgust - which were grouped in three categories: neutral, positive and negative.

We analyzed the .csv files by the timestamps and emotions, and combined the output data with the demographic information gathered from the screener survey. For each participant, we generated a report with their personalized charts analyzing their responses to each question, including a comparison with the other participants. These reports were shared with participants individually via follow-up sessions on Zoom, where they evaluated their results together with the investigators.

Target Outcomes

Our goal was to better understand how emotion-tracking software analyzes mock interviews and the factors that may influence preferences and attitudes towards various forms of feedback from the tool (content and formats).

While facial recognition adds complexity and stress in interview settings, emotion-tracking outputs can be used for increased self awareness in behavioral interviews. We hope to empower people interviewed with AI and encourage transparency and helpful feedback loops from AI interview-prep companies.

II. Literature Review

Langer et al. (2016) conducted an experiment to analyze the effect of a virtual employment interview (VI) training in candidates, with focus on nonverbal behavior. They used the software Visual Scene Maker and Microsoft's Kinect camera. In a simulated interview with a virtual character, a computer analyzed participants' nonverbal behavior (e.g. smiling, eye contact, posture, gesture, and voice characteristics) and provided real-time feedback. Results showed that participants of the VI training had a higher probability to receive a job offer. VI training reduced the anxiety of the candidates, helped them improve their nonverbal behavior, and seemed to be more effective than classic interview training. (Langer et al. 2016)

Suen, Chen & Lu (2019) investigated the synchrony effect by comparing human interviewer ratings and applicants' attitudes between asynchronous video interviews (AVIs) with AI decision agents and synchronous video interviews (SVIs) used in employment screening. The authors argue that candidates prefer human over AI rating, regardless of whether video interviews are synchronous or asynchronous. However they are less favorable towards asynchronous interviews because they must answer questions and watch themselves answering during non-human interactions, which are similar to Langer et al. (2017) findings.

From a pessimistic perspective, Harwell (2019) emphasizes *HireVue* could penalize nonnative speakers, visibly nervous interviewees, and those who do not fit in the AI model of look and speech. The opacity of this technology is another critique the author addresses to *HireVue*, as candidates do not know their score, what things they did wrong and how they could do better in the future.

In a recent study, Suen & Hung (2023) conducted a test to analyze job applicants' trust in AI asynchronous video interviews (AI-AVI) used in initial employment screening. Results show that applicant's trust AI-AVI more than in non-AI video interviews. Moreover, when the AI- AVI had features of tangibility and transparency, the applicants' cognitive and affective trust increased.

Suen & Hung (2023) recommend that [1] job applicants should be informed that their interview performance will be assessed by AI algorithms; [2] information about the AI model used and how it will assess the interviewees should be conveyed with simple words through text to increase cognitive trust by conveying transparency; [3] a virtual agent or avatar should be displayed on the screen to increase affective trust and simulate a human interaction in interviews.

Considering the company's perspective, the adoption of AI technologies is seen as a competitive advantage, which may reduce costs and time, selecting the best applicants that perfectly match the job offer. Van Esch and Black (2019) examined larger implications of the use of AI in hiring and proposed recommendations for companies grouped under "the three I's of AI-enabled recruiting: investigate, iterate, and integrate".

By investigating proactively [1], companies should adopt AI to screen candidates quickly and more effectively to decrease bias and increase candidate diversity. By iterating relentlessly [2], companies should employ different AI providers to test capabilities, in order to learn fast and less expensively. By integrating intelligently [3], companies should adopt chatbots to answer candidates quickly and fill in missing candidate information. (Van Esch & Black 2019, 736)

Recruiters and HR professionals generally do not perceive AI-enabled software as a threat, but as another tool that can simplify the search process, which can be an advantage in a highly competitive hiring space (Li et al., 2021). Besides believing that AI will continue to improve, some recruiters fear losing a strong candidate because AI could impact the

search and limit their options. Li et al. (2021) suggest the use of manual features together with the AI tools, in a hybrid approach. With a similar conclusion, Gonzalez et al. (2022) advocate for an augmented approach where AI/ML would be used in addition to human decision-making, and that applicants should be aware of the use of AI technologies. Gonzalez et al. (2022) suggest organizations reflect upon the context in which they are adopting AI/ML technologies in hiring processes, which methods are being used, how this impacts the applicant's performance, and the industry norms surrounding technology use.

Accounting for bias

Raghavan et al. (2020) uses algorithmic pre-employment assessment as a case study to show how formal definitions of fairness allow us to ask focused questions about the meaning of "fair" and "unbiased" models. The study identifies 18 vendors of algorithmic pre-employment assessments, documents what they have disclosed about their development and validation procedures, and evaluates their practices, focusing particularly on efforts to detect and mitigate bias. Raghavan et al. (2020) observes a heterogeneity in vendor's practices related to concerns about bias, which indicates they are sensitive with this topic but there is no clear guidance on how to respond to these worries.

This study shows that most of the vendors (15) offer customizable assessments, adapting their technologies to the client's particular data and job requirements. While 15 vendors made at least abstract references to "bias", only 7 explicitly discussed compliance or adverse impact of the assessments they offered. In particular, *HireVue* and *Pymetrics* described in detail their approaches to de-biasing their models, which involves removing features correlated with protected attributes when adverse impact is detected. (Raghavan et al. 2020)

Raghavan et al. (2020) emphasizes the context surrounding the use and deployment of a technical system, as design decisions should be analyzed based on relevant legal, historical, and social influences. They end their study with five policy recommendations: [1] Transparency is crucial to understand these systems; [2] Disparate impact is not the only indicator of bias, and vendors should also monitor other metrics like differential validity; [3] Outcome-based measures of bias are limited; [4] We may need to reconsider legal standards of validity under the Uniform Guidelines in light of machine learning; and [5] Algorithmic de-biasing techniques have significant implications as they automate the search for less discriminatory alternatives. (Raghavan et al. 2020)

III. Methodology Overview

Setting goals

To inform the research plan, we first needed to define our study goals as well as assumptions about what user goals might be when using AI emotion-tracking tools for interview practice. Our goals were to learn [1] What is the role of exhibited emotion in real video interviews? [2] What are human goals in using emotion-tracking software for mock interviews? [3] What are the factors influencing preferences and attitudes towards various forms of feedback from the tool (content and formats)? We formed base assumptions that users had the following goals:

- To see how well the interview went; how AI perceives their performance (visual display of emotion only)
- To see discrepancies between how they felt/thought they exhibited emotion vs. AI's interpretation
- To see what aspects of exhibited emotion they need to work on
- To understand what mix of emotions a person should exhibit or not exhibit during an interview; What does "good" look like?
- To achieve a "good" outcome in the tool so they feel more prepared for a real interview

Test planning

In addition to our initial project proposal, we documented a thorough testing plan to align on the details of our methodology and expectations. Upon review of the plan, which included a participant quota, draft screener survey, draft interview questions and learning goals from the experiment, we obtained instructor acceptance of the plan. Much of the plan's content is provided in the sections that follow.

IRB and human subjects training

In the interest of ethical research practices, team members completed the CITI Program course for social/behavioral research on human subjects that is usually required before requesting a review of the study proposal from the International Review Board (IRB). We did not submit our pilot study proposal to the IRB, however we did learn about code of ethics, federal regulations, informed consent, privacy and confidentiality. Although not required for our course or this pilot study, the IRB training helped us identify potential risks to our participants: [1] Sharing of personally identifiable information; [2] video recording of participants; [3] sharing of personal experiences/feeling vulnerable; [4] intentionally causing participants to experience anxiety during interview; and [5] possible feelings of anxiety during AI feedback portion.

TASK NAME	START DATE	END DATE	Jan	Feb				Mar				Apr			
			1	2	3	4	5	6	7	8	9	10	11	12	13
Preliminary research	1/23	2/9													
IRB ethics training	2/9	2/20													
Project proposal development	1/23	2/20													
Technology set up, design of benchmark data templates	2/6	3/8													
Designing the study	2/3	3/6													
Preparing mid-term presentation	2/25	3/6													
Generating benchmark data set	3/16	3/30													
Screener survey, recruitment	3/20	3/26													
Conducting interviews	3/27	4/3													
Testing algorithms	3/30	4/13													
AI analysis, data visualizations, reports generation	3/28	4/10													
Follow-up sessions, final data collection	4/7	4/14													
Statistical analysis, final presentation, report	4/10	4/24													

Table 1. Simplified version of the project roadmap and timeline focusing on milestones over 3 months.

Timeline and milestones

We established a roadmap with clear milestones over a 3 month period to keep the project on-track.

Identifying constraints and pre-mortem

We identified the following constraints: [1] We would be limited mostly to participants from our personal networks; [2] we needed to use an off-the-shelf AI tool that wouldn't require too much effort to make usable for the study; [3] we needed to recruit a generalized yet equitable group of participants that would not require special accommodations in order to take part in the study.

We also considered what might go wrong during the mock interviews: [1] Participant is not able to get camera or audio to work; [2] Participant is suddenly without internet connection; [3] Participant is not able/willing to answer the questions; [4] Interviewer is not able to access recording software or is unable to capture both video and audio; [5] Team is unable to retrieve the recording; [6] Team is unable to upload the recording to the application for analysis; [7] Recording file is corrupt; [8] Application freezes during analysis processing.

Defining target inputs and outputs

As inputs, we intended for participants to provide realistic verbal responses to interview questions via video call, with facial expressions and voice being recorded. As outputs, we intended for an AI analysis of facial expressions to generate a CSV file from which we could create and present at least 3 distinct visual formats for participants to respond to.

IV. Recruitment

Reasoning to inform participant quota strategy

We identified factors that we believed might influence the interest, performance, AI analysis, reactions, and preferences of participants:

- **English language proficiency:** All questions will be in English and our team requires responses to be in English so that we can assess what participants are saying against the tool's analysis of their expression at that moment
- **Sex at birth and gender identification:** Could this impact the person's display of emotion and/or how the tool recognizes the emotion displayed?
- **Age:** Does this impact the person's display of emotion and/or how the tool recognizes the emotion displayed? Does this impact a person's understanding of / level of comfort with AI tools?
- **Skin tone:** How does this impact the tool's ability to pick up facial features and expressions?
- **Ethnicity:** Are there any biased patterns in the tool's analysis of racial groups beyond skin tone?
- **Job Seeker Status:** Does this impact the level of motivation during the mock interview and the emotions displayed? Perhaps active job seekers are most interested in participating?
- **Familiarity with AI technology:** This may impact a person's expectations, performance, reactions, preferences, and attitudes towards the tool and the feedback provided
- **Experience with AI emotion analysis tools:** This may influence a person's attitude towards the tool

Quota and screener survey

The target was to recruit N=8 participants through a screener survey dispatched via the research team's social networks. Questions were designed to obtain a relevant and equitable sampling to meet the participant quota:

Fig 1. Summary of results from screener survey for participant recruitment.

- 12 questions on interest/availability, demographic information, job seeking status and familiarity with AI technology
- 32 respondents
- 5 did not pass screener = 27 viable respondents

Selecting an equitable sample

Below outlines the process we followed to clean the survey data and make our participant selections:

1. Reconciled for 1 survey question that changed after the survey was released
2. Added a column to summarize "Two or more race/ethnicity" (Yes/No)
3. Prioritized sampling by (i) age, (ii) skin tone, (iii) gender identification, (iv) familiarity with AI.
4. Shortlisted 12 participants with 4 alternates
    - 1 backed out
    - 2 no responses
    - 9 participants total were interviewed

V. Interview Sessions

Participant invitation and preparation

30-minute one on one mock interview sessions were scheduled with participants over one week. Participants were invited via Calendly and provided information about the nature of the session, the anticipated length of the session (15 minutes), and reminded to have [1] a reliable desktop computer set up in a quiet and private location, [2]

to have good lighting for the face, and [3] to have a Zoom account with reasonable familiarity on how to use Zoom.

Interviews

The research team scheduled nine remote one-on-one mock interview sessions over one week. Verbal informed consent to record was obtained at the start of each session. Interviewers followed a script and asked three behavior questions followed by three post-interview questions.

The team curated a list of three widespread behavioral interview questions that did not require domain knowledge to elicit a range of emotions from positive to negative:

1. Can you tell me about yourself?
2. Can you tell me about a time you went above and beyond?
3. Can you tell me about a time you overcame a team conflict or challenge?

The criteria for selecting interview questions were familiarity, universality, level of difficulty, clarity, and range of intended emotions evoked. We characterize familiarity and universality as how common the interview questions would be in a standard behavioral interview, regardless of the participant's domain. The questions were perceived by our team to contain an easy to moderate difficulty and be straightforward enough such that participants did not need to seek clarification from the interviewer. Furthermore, we wanted the emotions captured in the initial seconds of each response from the participants to closely represent their natural emotions and we wanted to avoid capturing any periods of uncertainty and confusion caused by the quality of questions selected.

The first question was intended to set an individual's emotional baseline. Since no two interviewers demonstrate the exact same mannerisms and possess the same technical expertise, this question allowed us to better analyze natural tendencies in each participant's expressions and communication style, specifically their natural level of positive, neutral, and negative emotions conveyed. This proved to be helpful when generating the reports and facilitating the feedback sessions when certain candidates demonstrated unintended emotions and did not agree with the results. The second question intended to showcase positive emotions related to participants' standout achievements in previous roles. The third question was intended to elicit negative emotions when recounting past events that required conflict resolution or overcoming significant challenges.

Following the mock interview, we asked participants (1) what emotions of the 7 they believed they displayed the most, (2) how the knowledge of AI-presence in the simulated mock interview impacted their interview performance or changed their prescription of their interview performance after the fact, and (3) if they had any prior experience with mock interview tools.

Evaluating model performance

To evaluate the model performance, after conducting the Facial Expression Recognition (FER) analysis, one technical team member analyzed each video from each participant by comparing the participants' facial expressions in the saved response videos and matching anchor points to the .csv file and line chart. We used the Happy, Neutral, and Sad+Fear emotions as a relative benchmark for positive, neutral, and negative emotions respectively. If the general trend of positive, neutral, and negative emotions were correlated with various snapshots in the videos, the results were considered valid. The context of the answers provided were not considered and the performance of the interviews as a whole were neither evaluated nor communicated with the participants in any capacity. Although not all participants agreed with the results of the emotion-tracking, all of the generated analytics by FER were deemed valid by the technical team member and there were no identified malfunctions that compromised the perceived functionality of the FER tool. Upon generating the feedback reports with visualizations, each respective interviewer verified the accuracy of the reports and analyzed the interview videos in order to explain any unintended analysis results in the feedback sessions.

Follow-up sessions

Participants were invited to return for remote one on one follow-up sessions to review their custom data visualization reports and share their responses to the AI-generated feedback. We presented each participant an interactive Figma prototype containing their report, which included data visualizations, video clips of their mock interview responses, and how their results compared to the results of other participants. We provided a contextual overview of the inspiration for the study (HireVue), and asked final questions. We wanted to know how the participants felt about their results, if they agreed with them, and if their feelings about AI facial emotion-tracking during interviews had changed from before. We also asked if they would ever use such an AI tool as part of their preparation for interviews. Participant responses were recorded in a spreadsheet to facilitate analysis.

VI. Technology

The AI-tool in this project was the Facial Expression Recognition Python Library (Shenk et al., 2021) which enabled us to conduct the mock video interviews over Zoom and analyze the emotion tracking results asynchronously. We used the Multi-Task Convolutional Neural Network (MTCNN) option for face detection and the default Keras API for the backend. The FER model was trained on the FER-2013 dataset consisting of 28,709 images of 48x48 pixel grayscale images of faces categorized into 7 emotions: angry, disgust, fear, happy, neutral, sad, and surprise. The dataset contained images that

obtained multiple classifications but most training images were dominated by exaggerated expressions. Many of the classifications such as anger+disgust and surprise+fear were similar in expression and distinguished primarily by eye shape and width.

Each video was analyzed frame by frame to classify the emotions into seven categories: neutral, happy, sad, fear, surprise, happy or disgust on a scale from 0 to 1. These results were compiled with timesteps of the video into a csv format for further analysis and visualizations.

VII. Analysis & Visualization

To gain insights from the participant's mock interview and the intensity of the emotions for each question, we plot a line chart of each participant's raw emotion output against time for each question. As we can see from the figure below, it is difficult to distinguish between the seven emotions in the visualization.

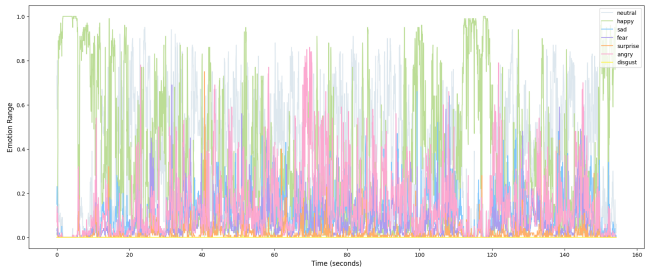

Fig 2. Emotion chart for participant #2, question 2. The chart visualizes the 7 emotions tracked over time.

We wanted the participants to easily discern the emotions and seek to simplify the visual representations of their emotions. Hence, we merged the output data for all participants and questions, and performed a statistical analysis to observe if a trend emerges.

	Candidate Number	Question	Time (mm:ss)	Time (sec)	neutral	happy	sad	fear	surprise	angry	disgust
0	1	1	0:0.0	0.00	0.63	0.16	0.15	0.01	0.00	0.05	0.0
1	1	1	0:0.04	0.04	0.57	0.22	0.15	0.02	0.00	0.04	0.0
2	1	1	0:0.08	0.08	0.56	0.26	0.13	0.01	0.00	0.04	0.0
3	1	1	0:0.12	0.12	0.56	0.26	0.13	0.01	0.00	0.04	0.0
4	1	1	0:0.16	0.16	0.56	0.27	0.13	0.01	0.00	0.04	0.0
...	...	...	...	...	...	...	...	...	...	...	...
1880	9	2	1:15.2	15.20	0.77	0.06	0.11	0.02	0.00	0.04	0.0
1881	9	2	1:15.24	15.24	0.59	0.16	0.17	0.02	0.01	0.05	0.0
1882	9	2	1:15.28	15.28	0.56	0.17	0.17	0.04	0.02	0.05	0.0
1883	9	2	1:15.32	15.32	0.55	0.17	0.16	0.05	0.03	0.03	0.0
1884	9	2	1:15.36	15.36	0.70	0.15	0.11	0.01	0.01	0.02	0.0

Table 2. Merged output data from the model.

	neutral	happy	sad	fear	surprise	angry	disgust
count	51549.000000	51549.000000	51549.000000	51549.000000	51549.000000	51549.000000	51549.000000
mean	0.348248	0.254941	0.135430	0.124950	0.047705	0.080123	0.007631
std	0.257092	0.319414	0.137797	0.138469	0.105166	0.091537	0.039976
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.120000	0.030000	0.030000	0.020000	0.000000	0.020000	0.000000
50%	0.320000	0.100000	0.090000	0.080000	0.010000	0.050000	0.000000
75%	0.550000	0.360000	0.190000	0.180000	0.040000	0.110000	0.000000
max	0.990000	1.000000	0.880000	0.890000	0.910000	0.870000	0.870000

Table 3. Statistical analysis of each emotion in the dataset.

Through the statistical analysis above, we observed that neutral has a distinct distribution, whereas happy and surprise has a similar pattern. The distribution of sad, angry, and fear are quite similar, while disgust is observed to be an outlier between the seven emotions. Hence, we decided to group happy and surprise as positive, while sad, fear, angry, and disgust are grouped as negative, with neutral remaining as-is. When grouping the emotions, we maintained the data integrity by keeping only the maximum value observed of each emotion group per row. Then, we created a summary graph of the three emotions against time as seen in the figure below.

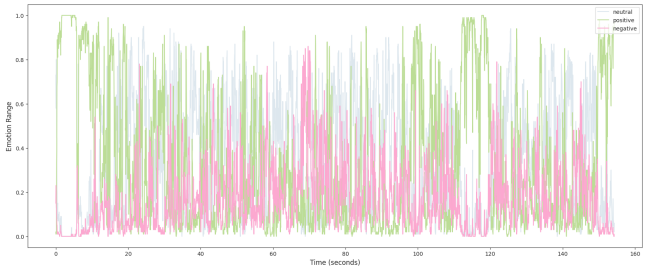

Fig 3. Summary chart for participant #2, question 2. The 7 emotions are summarized to 3 distinct emotions: positive, negative and neutral.

For the report prototype, we created individual bar charts of emotion frequency and emotion amplitude for each participant per question to give a detailed view on how their positive, negative, and neutral emotions differ for each question. At the same time, we also provide a summary of their overall emotion frequency and emotion amplitude as a bar plot, which are then displayed in conjunction with other participants' results to make it easier for them to view the difference. In this instance, emotion frequency is the normalized count of top emotion type observed for each record, and emotion amplitude is the mean value of the top emotion observed for each record.

Next, we pre-processed the screener survey results and looked into participants' profiles to segment our participants into distinct groups and subsequently create visualizations to observe trends between the groups. From their profile, we opted to segment according to demographic information on age group, gender identity, AI familiarity and skin tone.

Based on the participants profile, the breakdown for each

demographic category is as follows:

- *Age group:* 3 participants in 18-24, 2 participant in 25-34, 3 participants in 35-44, and 1 participant in 45-54 age range
- *Gender:* 6 Female and 3 Male
- AI familiarity: 2 casual user, 2 hobbyist, 3 familiar and 2 advanced
- *Skintone:* 3 participants for fair to light, 2 participants for light to medium, and 3 participants for medium to dark, and 1 participant for dark to very dark

To create the demographic comparison barplot, we first merged the participants demographic information with their emotion data. Then, further data processing are done at question and individual level prior to generating graphs.

In the resulting charts, we observed that participants in the 18-24 age group are more likely to display positive emotions compared to other age groups. The gender comparison shows expected results with male displaying mostly neutral emotion compared to women. However, we are not able to conclude a unifying trend in terms of AI familiarity as there does not seem to be a strong correlation between the emotions observed and the participant's familiarity with AI. Finally, the skin tone comparison chart shows that the FER model was able to recognize emotions across different skin tones effectively.

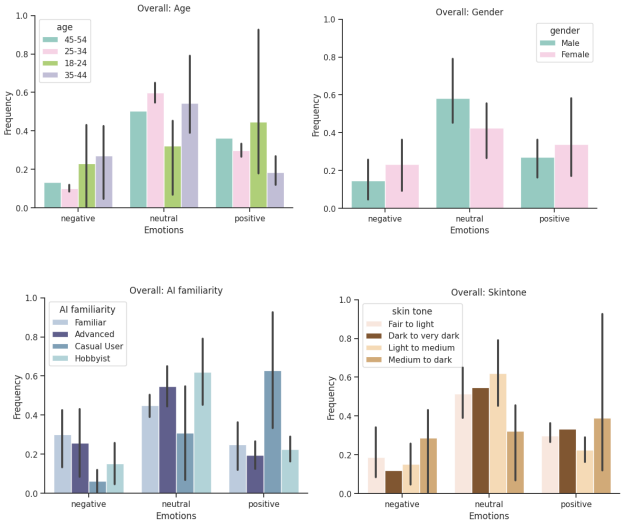

Fig 4. Bar plot of demographic comparison based on emotion frequency. From top left to bottom right: Comparison based on age, gender, AI familiarity, and skin tone. The bar plot uses a range between 0 to 1 and illustrates the overall trend for negative, neutral, and positive emotions.

Aside from the bar charts displayed above, we also produced demographic comparisons at question-level, displaying the variability of emotion frequency of each demographic group for each question. The bar chart mentioned can be viewed on the report prototype.

VIII. Discussion (Follow-Up Insights)

The target of the follow up insights were to address the similarities between how the candidate thought they performed and how the AI tool analyzed their results. By performance, we refer to which emotions the participants think they conveyed the most. Many of the participants listed happy and neutral as their primary emotions and those who were most surprised by the AI results were presented with higher negative emotions than they anticipated.

As part of the discussion sessions, we replayed select interview answers and then reviewed the facial recognition results again. Most candidates who previously agreed with the results became more confident in the validity of the results after seeing their previous interview responses. Similarly, candidates who initially disagreed with the results began to question how their natural resting facial positions, webcam quality, camera angle, lighting, and background environment predisposed them to reading higher negative emotions. For most interviewees, especially college students, it is difficult to change these peripheral details while consciously modifying one's natural face expression can be a cause of stress and frustration among candidates. For another candidate, they rejected the results altogether which brings bias into the discussion when using a single AI model to evaluate one's emotions.

Concerns

Many participants felt AI would have an increased presence in interviews going forward, and many were not aware that employers have been using AI to analyze emotions for years. Participants anticipated that they will have to manage their expressions and mannerisms both for in-person interviews and those conducted virtually/remotely.

Satisfaction

Regardless of whether people were prepared for the interview or actively looking for work, 7/9 participants found value in using an AI tool for emotional analysis as part of virtual interview prep.

External factors

We did not anticipate the lack of preparedness for mock interviews. We suspected participants wouldn't equate this experiment with a real interview platform, so we opted to sculpt the experiment around a mock interview preparation tool. However, we found that participants were even unprepared for our mock interview even if they were actively looking for work outside of our experiment. We had more than one issue of participants taking interviews outside or even while walking their dogs which they would never do in a real mock interview setting.

Participants were hand picked within our social network which made it seem more like a personal favor for a school project rather than a formal research project.

Discussions

Not all participants agreed with their results, and they were forced to wonder whether it was their natural expressions or interview environment that could have made the AI gather those results. For those affected, it was a disappointing result, but they still found the experience to be insightful and interesting.

Others decided to push back: respecting the impact that facial recognition could have for the hiring processes going forward but also refusing to give in. While most people simply cannot make their own business, and while most startups fail, it is interesting to see how much the use of AI in interviews will deter future applicants from applying (like with excessive take home programming assignments).

IX. Future Work

Bias studies

If emotional analysis is used in interviews or in a practice interview tool, which skin tones, demographics, income brackets, age groups, and genders are most disadvantaged to be perceived well by the AI in such interviews?

Do better peripherals (camera quality, camera angle and distance, artificial lighting, access to natural light, arbitrary bookshelves in backgrounds, etc.) matter?

Do the AI results regarding responses to behavioral questions matter in more technical fields where behavioral skills are discounted (e.g. software developers, engineers, etc.) compared to fields where strong behavioral skills are required (consulting, banking, law, counselors, teachers, etc.)?

Hiring Companies

Because this study focused on how people using an AI-based practice interview tool feel about the process and the results, we did not explore how emotional analysis through facial expressions translates to predicted characteristics and perceived interview performance by a corporation. It is common for medium and large corporations to have defined hiring guidelines and evaluation criteria for potential candidates; however, there is great variability from industry to industry, corporation to corporation, and interviewer to interviewer. In future work, it could be insightful to learn how hiring companies or intermediary staffing companies that use AI vs. don't use AI perceive the results of this study and what impact that may have on their own interview evaluation criteria.

Natural language processing

Our study did not attempt to evaluate any participant's performance or establish a sweeping evaluation criteria for candidates who all had different personalities, backgrounds, genders, levels of experience, and motivations for participation in our study. As stated in our introduction, we do not believe emotion-tracking results can accurately correlate to personality traits or a fair evaluation of a person's suitability for a job position. However, we did come up with some general insights for anyone being evaluated by AI in an interview setting, mock or real.

In general, when answering behavioral questions, it is advisable to put a positive spin on answers while clearly outlining the context, solution, and impact of the story. Our model was able to detect negative facial expressions but future studies involving natural language processing could explore whether negative facial expressions are perceived similar to negative contexts. Such a study would better inform interviewees on how much impact "what you say" has on interview performance relative to "how you say it."

X. Conclusion

This study's results are not meant to or sufficient to affirm if AI systems are able to accurately measure emotions, however the analysis indicates opacity about which facial expressions are labeled. Not all participants agreed with their results, and they were forced to wonder whether it was their natural expressions or interview environment that could have made the AI generate those results.

Candidates who agreed with the results became more confident after seeing the AI report. Similarly, candidates who disagreed with the results questioned if facial position, webcam quality, lighting and background environment interfered with AI analysis, which resulted in more negative emotions.

We had participants with different skin tones, from light to very dark, and all interviews were successfully analyzed by the FER model. Despite these results, we cannot evaluate if the skin tone interferes with emotion tracking and further studies regarding bias should be addressed.

There is a large incentive for interview candidates to prepare for AI tools in conjunction with typical job preparation activities (such as updating resumes, company research, applying for jobs, mock interviews, and networking). For instance, there are AI tools that exist for building resumes that pass AI-driven Applicant Tracking Systems (ATSs).

There are also AI tools that help people practice for interviews with warm-up questions. To our knowledge, while most candidates are aware that facial recognition

tools are in use, there is less common knowledge regarding emotion tracking outside of the tech sector and there is far less common knowledge regarding using AI emotion tracking as a tool to improve performance in virtual interviews.

REFERENCES

Chakraborty, Ishita, Khai Chiong, Howard Dover, and K. Sudhir. "AI and AI-Human Based Salesforce Hiring Using Interview Videos." Available at SSRN 4137872 (2022).

Gonzalez, Manuel F., Weiwei Liu, Lei Shirase, David L. Tomczak, Carmen E. Lobbe, Richard Justenhoven, and Nicholas R. Martin. "Allying with AI? Reactions toward human-based, AI/ML-based, and augmented hiring processes." Computers in Human Behavior 130 (2022): 107179.

Harwell, Drew. "A face-scanning algorithm increasingly decides whether you deserve the job." In Ethics of Data and Analytics, pp. 206-211. Auerbach Publications, 2019.

Langer, Markus, Cornelius J. König, Patrick Gebhard, and Elisabeth André. "Dear computer, teach me manners: Testing virtual employment interview training." International Journal of Selection and Assessment 24, no. 4 (2016): 312-323.

Langer, Markus, Cornelius J. König, and Kevin Krause. "Examining digital interviews for personnel selection: Applicant reactions and interviewer ratings." International journal of selection and assessment 25, no. 4 (2017): 371-382.

Li, Lan, Tina Lassiter, Joohee Oh, and Min Kyung Lee. "Algorithmic hiring in practice: Recruiter and HR Professional's perspectives on AI use in hiring." In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 166-176. 2021.

Maurer, Roy. “HireVue Discontinues Facial Analysis Screening.” SHRM, February 3, 2021. https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/hirevue-discontinues-facial-analysis-screening.aspx.

Perkowitz, Sidney. “The Bias in the Machine: Facial Recognition Technology and Racial Disparities.” MIT Case Studies in Social and Ethical Responsibilities of Computing, February 5, 2021. https://doi.org/10.21428/2c646de5.62272586.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. "Mitigating bias in algorithmic hiring: Evaluating claims and practices." In Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 469-481. 2020.

Singh, Rahulraj. “The Ultimate Guide to Emotion Recognition from Facial Expressions Using Python.” Medium. Towards Data Science, July 26, 2021. https://towardsdatascience.com/the-ultimate-guide-to-emotion-recognition-from-facial-expressions-using-python-64e58d4324ff.

Shenk, Justin. “Fer.” PyPI. Accessed February 20, 2023. https://pypi.org/project/fer/.

Shenk, Justin, Aaron Cg, Octavio Arriaga, and Owlwasrowk. Justinshenk/Fer: Zenodo (version zenodo). Zenodo, 2021. https://doi.org/10.5281/zenodo.5362356.

Suen, Hung-Yue, Mavis Yi-Ching Chen, and Shih-Hao Lu. "Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes?." Computers in Human Behavior 98 (2019): 93-101.

Suen, Hung-Yue, and Kuo-En Hung. "Building trust in automatic video interviews using various AI interfaces: Tangibility, immediacy, and transparency." Computers in Human Behavior (2023): 107713.

Van Esch, Patrick, and J. Stewart Black. "Factors that influence new generation candidates to engage with and complete digital, AI-enabled recruiting." Business Horizons 62, no. 6 (2019): 729-739.