Towards Building Personalized Facial Emotion Recognition Systems

Hailey Park
hpark353@wisc.edu
Statistics Department, University of Wisconsin-Madison
Madison, WI, USA

Michael F. Xu
michael.xu@wisc.edu
Computer Sciences Department, University of Wisconsin-Madison
Madison, WI, USA

Naihao Xu
Computer Sciences Department, University of Wisconsin-Madison
Madison, WI, USA

Matthew Bayles
Computer Sciences Department, University of Wisconsin-Madison
Madison, WI, USA

ABSTRACT

Facial Emotion Recognition (FER) systems are becoming more pervasive with the advent of more efficient algorithms and more powerful edge computing devices. While the average performance of such systems are impressively high, accuracy on various demographic subgroups or specific individuals may vary a lot, where factors as age, gender, and ethnicity are all important determinants. We posit that in many scenarios, such as in long-term interactions, there is untapped potential in personalizing such FER systems. In this project, we explore the first step in that direction: understanding users' needs and preferences in terms providing feedback to help the system learn over time. We developed an online questionnaire, and distributed it to 38 participants. Among other insights, we found that users preferences significantly depends on the deployment context of the FER system. We provide design recommendations based on these findings.

KEYWORDS

Facial emotion recognition, Personalization, Human-Computer Interaction, User feedback

ACM Reference Format:
Hailey Park, Michael F. Xu, Naihao Xu, and Matthew Bayles. 2024. Towards Building Personalized Facial Emotion Recognition Systems. In . ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Facial Emotion Recognition (FER) systems have become deeply embedded in various aspects of our lives, as researchers and industry practitioners have found viable applications for such models in sectors ranging from healthcare [6], education [21], marketing [8], and beyond. The latest FER systems are generally able to achieve high accuracy levels when evaluated on the aggregate performances [15], and the addition of FER to existing systems and platforms have demonstrated real, practical values to their adopters. While the aggregate performance are high, accuracy on various demographic subgroups often vary much more, where age, gender, and ethnicity have all been shown to be significant factors [4, 12, 16]. Researchers have attempted to address such concerns from both the dataset, and the algorithmic perspective [5], yet most of such studies focus on the application of an FER system that is meant to face a large and random user group.

While FER systems are increasingly deployed to personal or domestic domains, there is a surprisingly lack of literature investigating FER systems intended for long-term usage by one or a few specific users. We posit that, over repeated interactions with the same user, the systems would have the opportunity to improve itself, and evolve into a *Personalized Facial Emotion Recognition* **(PFER)** system that could achieve more accurate emotion readings on the specific user it serves. In this paper, we explore one of the first steps towards understanding how one might build such a PFER system. Through an exploratory online questionnaire, we've described hypothetical settings and scenarios, and elicited participants' responses on their perceptions and preferences regarding FER systems, a personalized version of it, and interactions with such a system. Specifically, our research questions are as follows:

(1) **RQ 1:** What are users' perceptions about such a PFER system?
(2) **RQ 2:** What is the preferred channel and mechanism for users to provide feedback to a PFER system?
(3) **RQ 3:** How does the preference depend on the deployment context of the PFER system?

2 BACKGROUND

To the best of our knowledge, human feedback solicitation in the context of personalized FER systems has not been explicitly studied before. Nevertheless, prior work in related fields of human-AI interaction, and applications concerning general feedback solicitation from human, provide us with insights into the potential dimensions one may explore in the context of personalizing FER systems.

2.1 Soliciting Human Feedback

While there have been a large number of studies concerning human-in-the-loop machine learning systems, almost all of them focus on

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

the effects of the human feedback on the system's performance, where the human's role and attention was dedicated to providing annotations and feedback to the system [9, 18, 20]. A few others have studied how the user's perception of the systems may change as a result of providing feedback, but all have treated the feedback mechanism as a static constant, where again the human's attention was dedicated to providing the feedback [7, 23]. In a naturalistic, everyday setting, however, instead of situated within a turn-taking environment, human-AI interactions are often much more fluid and continuous, with the user juggling multiple tasks at once, and only interacting with the AI system as one of the competing priorities. In these more realistic scenarios, *how* should the system solicit human feedback?

Some recent work in the field of Human-Robot Interaction (HRI) started exploring these topics. In a fast-paced, simulated collaborative task between a human and a robot, Candon et al. [3] studied the effect of the framing and timing of the reminders for feedback. The authors found that the timing of the reminders had a significant impact on the frequency of feedback received from the human teammates, and while framing the reminder as a means to improve team performance vs. system performance did not impact the feedback frequency, it did improve the users' perception about the robot, as well as feeling more positive about providing feedback to the robots. In a study where the authors investigated potential methods to align implicit and explicit human feedback, Zhang et al. [24] noted that explicit feedback such as directly asking users for a demonstration, for their preference, or for their evaluative feedback, can interrupt the natural flow of the interactions. Coincidentally, one of the modalities of the *implicit human feedback* discussed was through analyzing users' facial expressions, highlighting the value of a PFER system where a system could gradually shift from an explicit feedback mode to implicit feedback mode without loss of accuracy.

2.2 Human-AI Interactions

Several important factors that may impact the interactions between the user and a system have also been identified within the broader context of human-AI interaction. Van Berkel et al. [22] described the three paradigms of interaction as intermittent, continuous, and proactive. A PFER system soliciting feedback would fall into the continuous interaction paradigm, as it continuously reads in user input (i.e. facial expressions). Reaffirming the description of the more realistic scenarios in the previous section, the authors suggest that, for such interactions, users are often focused on a task, and distracting them from that activity could be undesired. It is therefore potentially important for a system to be able to interject as subtly as possible, and recognizing the best timing for these interruptions. Through an online user study, Liao and Sundar [14] found that the presumed role of the AI system also impacts user's perception and evaluation of the system, and that the effect is differential across different types of users. Specifically, the authors concluded that while power users exhibit higher trust when the system assumed the role of a help-seeker, non-power users are less concerned about privacy when the system indicates that it is both a help-seeker and a help-provider (i.e. it needs the user's help to better assist the user). The implication seems to be that the framing of the system matters, but that how it matters may be dependent on the user. More broadly, Li et al. [13] assessed some of the existing Human-AI interaction guidelines, and confirmed the importance of having an option for the users to provide feedback, and for the system to learn from such feedback. The authors, however, noted that these features maybe interpreted by users as a *promise* for improved system performance, and that long-term user perceptions may be different depending on how the system delivers on such implicit promises.

3 METHODS

We aim to understand how we can design personalized FER system through user feedback. Specifically, we study whether gathering more user-specific data through the deployment of the system can improve the accuracy of the FER model. Furthermore, we study how we can collect this data and user+ feedback without disrupting the original purpose of the system. We will employ two methodologies to investigate each aspects of personalized FER system design. First, we conduct an experiment that empirically tests whether gathering user-specific data can improve the model performance. Then, through a web-based questionnaire, we investigate user preferences in terms of feedback collection mechanism employed by such systems.

3.1 "Personalized" FER Model through User Feedback

How can user-specific data be incorporated in the existing FER model? We investigate effective ways to prompt user feedback. We postulate that collecting user-specific data in the form of user feedback can be effective in improving the FER model through adding personalizing touch to a pre-trained model. A FER system based on Deep Convolutional Neural Network (DCNN) architecture and Transfer Learning (TL) can be implemented to incorporate such user feedback. DCNN models achieve the state-of-the art accuracy in facial emotion reognition tasks [11, 17]. Additionally, transfer learning framework enables us to augment the pre-trained model with additional pictures of a user. Some studies have shown that transfer learning can also enhance the performance of the pre-trained network [19]. Below, we set out some steps we recommend future researchers to take in incorporating user feedback into an existing FER model.

3.1.1 System Design. Transfer Learning technique can be utilized to integrate user-specific images into the pre-trained dataset. Future research can emulate the model architecture from a prior study that employed transfer learning in FER model, such as [1]. We deem this system appropriate because the model's facial emotion recognition task aligns with our primary focus and the observed high model performance. This system can be expanded by crafting a dataset comprising facial images of an individual. This aims to replicate the hypothetical FER system product that collects user images while the system is in active use.

3.1.2 System Evaluation. In evaluating a system before and after applying the transfer learning, model performance can be compared before and after incorporating user feedback. Each participant can sit in front of a computer equipped with a webcam and a program with a basic user interface (UI) designed by the research team. The

researcher conducting the session will start the program and sit off to the side. The session can be done in person or virtually. The UI will display pictures and prompt the user to select which emotion they are experiencing while viewing the image. If the session is virtual the researcher will select the emotion for the participant. Example emotions are anger, fear, disgust, happiness, and sadness. A group of two sets of sixteen images can be selected by the research team. One set for each stage of the experiment. Each emotion being studied can have four images associated with it. The images can be shown in a random order with every image being shown exactly one time. The response of the participant and the response of the model can be recorded by the program. While the participant is being shown the images, the model captures an image of the participant experiencing emotion using the webcam for each image they see, if the session is in person. If the session is virtual the researcher will record the session and the participant images will be extracted at later data. These images can be used to predict the participants emotion and will be used in the transfer model. Using the user's input, the taken image and their response will be applied to the model using Transfer Learning. Once the whole series of images and responses have been captured, we can run the same procedure as above but with the updated model and a different set of sixteen images. Depending on the training time of the model and whether the session is virtual this can either occur in the same session or later.

3.2 User Preferences on Feedback Mechanism

Studies in human-AI interaction have demonstrated that the field is complex and multi-dimensional. Among other things, the role the system assumes [14], the proactiveness of the system [22], and the modality of the feedback [10], are all important considerations. For this project, our goal is to understand the distribution of user preferences over these various dimensions, specifically in the context of human feedback for FER systems.

3.2.1 Study Setting and Population. As there is limited existing research on this specific topic, we conducted an exploratory questionnaire with a total of 38 adult participants, consisting of friends and families of the research team. This population is generally tech-savvy and familiar with the various applications of systems involving digital faces, such as Apple's Face ID and conference calls, but may not be entirely familiar with the FER system itself (average self-reported familiarity with FER systems is around 25 on a scale of 0 to 100). More characteristics of the participants are shown in Table 1.

While our research question is generally applicable to scenarios where FER systems are intimately positioned along with human users, for this specific exploratory questionnaire, to help participants contextualize our questions and their responses, we will describe two specific settings involving FER systems: 1. An in-home social robot equipped with a camera for vision tasks, and 2. A screen-based application such as conference call software. Participants will be randomly asked about their preferences in one of these settings. For ease of comparison, the sets of questions for each scenario is essentially equivalent, except for some terms referring to the robot vs. the application.

Table 1: Demographic Information of Participants

	Frequency	Percentage
Gender		
Male	19	50.0
Female	19	50.0
Age Range		
16 - 20	2	5.26
21 - 25	12	31.58
26 - 30	3	7.89
Over 30	21	55.26
Education		
High School	2	5.26
Some College	6	15.79
Bachelor's	15	39.47
Master's	13	34.21
Doctoral	2	5.26

3.2.2 Study Design. In order to answer the research questions, especially *RQ 3*, we employed a between-subject design, where participants are randomly assigned one of the two deployment scenarios (robot *vs.* screen-based apps). To better control for individual-level variances, it would have been preferable to conduct a within-subject study, with the order of the scenarios counter balanced over the participants. However, our internal testing as well as pilot testers revealed that having participants complete both scenarios would have been too much of a burden, which could negatively impact the quality of data we receive. For these reasons, we've opted for the between-subject design. Along with this choice, we also decided that instead of the original target of 20 participants, we will aim to recruit 40 participants to increase the power of the analysis.

3.2.3 Data Collection. The questionnaire was created based on existing literature in related fields, and refined through workshop sessions among the research team. We asked participants to evaluate both high-level preferences such as modality and frequency of feedback solicitation, and low-level preferences such as comparisons of several variations of a feedback mechanism. We also included open-ended questions to allow participants to mention other aspects they deem important, but not covered in the questionnaire, and to elaborate on their responses. After several brainstorming sessions within the research team, we settled on a finalized version of the questionnaire, digitalized and deployed it on a web-based survey platform, Qualtrics. Links to the anonymous questionnaire were distributed to participants, who then completed the questionnaire on Qualtrics, where the data are collected and stored.

3.2.4 Questionnaire Structure and Sample Questions. The questionnaire starts with a set of standard demographic questions, where we also asked participants about their familiarity with FER systems. The participants are then randomly assigned one of the two scenarios, an FER system embedded on a domestic robot, and an FER system running on a standard screen-based application (e.g. conference app). In each case, the participants are first introduced to the scenario, where potential utility and use cases of FER are described to the users. Next, participants are asked to rate a series of questions

on sliders, with a scale from 0 to 100. An example question is "*In terms of providing feedback to the software's emotion readings, do you prefer that the software asks for feedback in a synchronous or asynchronous fashion?*" where we also provided a description of what each option would be reflected in the scenario. Another example is "If the software needs to ask for feedback, how do you prefer that it's initiated?" For this question, we included three options, each paired with its own slider for the participants to express their preferences. The three options are *Audio: Verbally asks through natural language*, *Visual: Text on display screen*, and *Visual & Audio: Text on display screen & a chime / "ding" sound.* Additionally, we've also touched on aspects such as the frequency of interaction, and preferences regarding ways to improve or customize such interactions. Throughout the questionnaire, we also included opportunities for open responses, and encourage participants to provide more context or suggest other interaction modalities or considerations.

3.2.5 Overview of Data Analysis. We will perform both qualitative and quantitative analysis on the survey responses. We will generate descriptive statistics and analyze for patterns for the quantitative sections. For the qualitative section, we plan to conduct a Thematic Analysis [2]. We will first generate potential codes, and then individually assign codes to the relevant responses. We will then discuss and iterate on the code assignment until an agreement is reached for each response.

4 RESULTS AND DISCUSSIONS

As for the quantitative responses, we ran a linear regression analysis to investigate how feedback modality preferences or feature preferences vary by different independent variables. The underlying assumptions for the linear regression models were met upon checking the following: normality, independence, and constant variance. We measured the goodness of the fit of the models by running chi-square tests for the nested models.

4.1 Feedback Modality Preferences

In analyzing the feedback modality preferences, we found that the user preference significantly varies by the FER model deployment scenario. In a screen-based application, textual prompting of user feedback was preferred in comparison to the audio prompting style in an additive model($\beta_{scenario} = -49.04, T_{scenario} = -5.50, p_{scenario} < 0$). Furthermore, users within the age range of 21-25 preferred the text prompting style more than those within the age range of 16 - 20 ($age_{21-25} = -53.82, T = -2.65, p < 0$). Likewise, users within the age range of 26-30 ($age_{26-30} = -46.51, T = -1.99, p < 0.06$) as well as users with age over 30 ($age_{over30} = -58.07, T = -3.01, p < 0.001$) preferred the text prompting style compared to the users who are 16-20. A model fit on the audio feedback prompting preference aligned with these findings from the textual feedback model. Users found the audio-based feedback prompting less attractive in a screen-based application compared to the robotic application ($\beta_{scenario} = -29.17, T = -2.93, p < 0.001$). Interestingly, there were no variation observed in the user preference between two deployment context in adopting both the visual and the audio feedback prompting. A concrete example of this prompting style is to display a text asking for user feedback on a screen along with a chime or a "ding" sound.

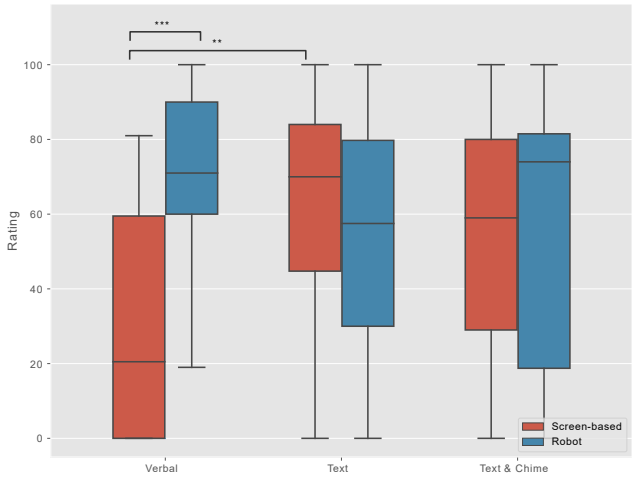

Figure 1: User Preferences: System Prompting for Feedback

We illustrate these findings in Figure 1, where we've performed both paired and unpaired t-tests, while applying the Holm-Bonferroni corrections for multiple tests. Consistent with our linear regressions, deployment scenarios significantly impacted user preferences for how the system shall prompt for a feedback. In addition, we also notice that within screen-based applications, user preference for text based prompting is higher than that for verbal.

We also analyzed whether there is any variation in how users prefer to provide feedback to the system after their feedback is prompted. Similar to the pattern observed in the feedback elicitation preferences, we found that users prefer to provide feedback in a textual format in the screen-based application compared to those using robotic applications ($\beta_{scenario} = -29.17, T = -2.93, p < 0.001$). This is also illustrated in Figure 2. What we found to be different from the modality of prompting feedback was that there was no variation observed in user preferences across different age.

4.2 Feedback Elicitation Frequencies

We asked user preferences for how often they would like to be prompted for user feedback in two cases: the initial case as users are getting used to the system and the system is beginning to learn about the granular expressions of the user and the familiar case after users have interacted with the system a few times and the system is more familiar with the user's facial expressions. Figure 3 illustrates the bar plot of user preferences in feedback prompting frequencies for the initial case. As can be seen from the figure, users initial prefer to be prompted for feedback at a moderate frequency: the highest votes in "Every few hours", "Every minutes", to "Whenever the system detects an emotion". This implies that users are willing to provide feedback at the initial stage as they are beginning to use the system to help the system make more accurate predictions and in making the model more personalized. Figure 4 shows the corresponding plot in the familiar case. The figure shows that the "Every few hours" still is the most popular option, although more frequent prompting options like "Every few minutes" and

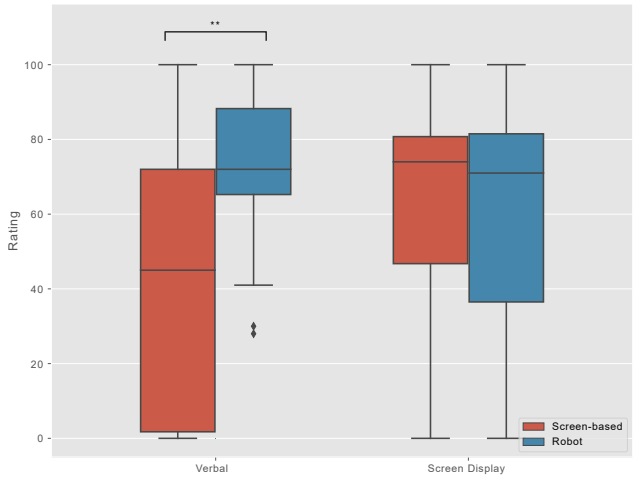

Figure 2: User Preferences: Providing Feedback to System

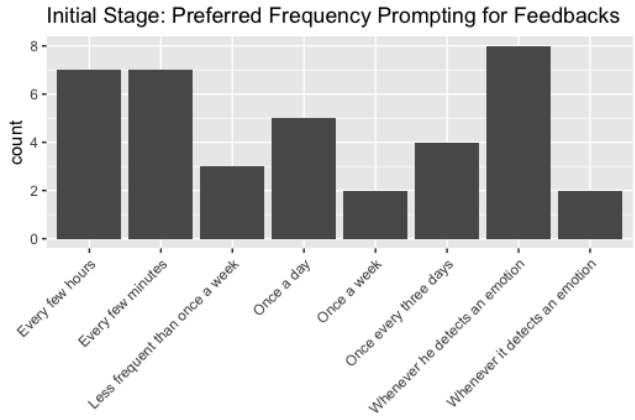

Figure 3: User Preferences: Initial Feedback Prompting Frequencies

"Whenever the system detects an emotion" has gathered less votes. These indicate that users in the later stages prefer to be prompted with user feedback at less frequency than the initial case.

4.3 Feedback Feature Preferences

Now, we consider whether there is variation across different users' preferences in features relevant to providing user feedback. First, we asked user opinions on a snooze feature that can disable the feedback eliciting for a while. A lot of the users found this feature very useful, rating this feature with a mean of 76.56 out of 100. There is some evidence that users within the age range of 26-30 ($age_{26-30} = 33.33, T = 1.73, p < 0.1$) and those over 30 ($age_{over30} = 27.67, T = 1.75, p < 0.1$) consider this more useful than users who are 16-20. There were no variations across deployment contexts nor gender.

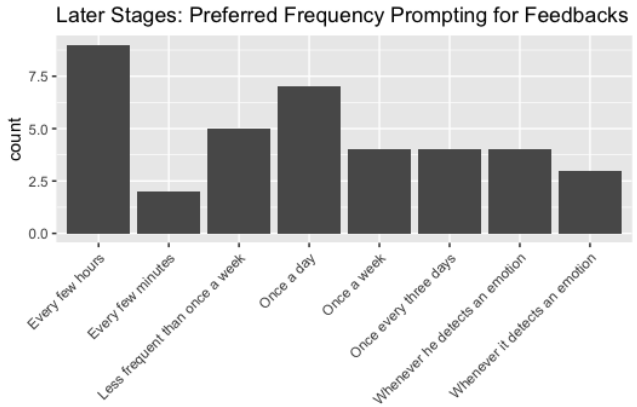

Figure 4: User Preferences: Later Sessions' Feedback Prompting Frequencies

Another feature we suggested was a switch that can completely turn off the feedback prompting. This had a mean rating of 69.82, receiving a moderately positive response but not as high in numerical rating compared to the temporary turn-off trigger. No significant variation was seen regarding this feature across deployment contexts, user gender, and age. There is some evidence that users who have received a master's degree find this less useful than those receiving bachelor's degree ($edu_{master's} = -22.17, T = -2.01, p < 0.1$) and that users who have obtained some college degree find this slightly more useful than those receiving bachelor's degree ($edu_{college} = 22.76, T = 1.75, p < 0.1$).

A final feature that we suggested to users was the user feedback being prompted at the beginning of every interaction whether they would like to provide feedback in this session or not. This feature also received a positive response with a mean rating of 75.61. Users found this feature to be a lot more useful in the web-based applications than in the robotic applications ($\beta_{scenario} = 24.61, T = 2.44, p < 0.05$).

We summarize some of these findings in Figure 5. Although some of the tests loses their significance after correcting for multiple-tests, the high-level picture is quite clearly illustrated here, where we find a strong dislike for the session-based permission-asking feature for the robotic deployment scenario.

4.4 Feedback Timing Preferences

We asked users to rate their preferences for synchronous feedback vs. asynchronous feedback. Result is shown in Figure 6. While we did not find a statistically significant difference between the two deployment scenarios, on average users demonstrated a stronger preference for real-time feedback interactions in the robotic scenario, while slightly leaning towards an asynchronous philosophy when it comes to a screen-based application. This may be reflective of the type of tasks and context that the users are envisioning happening in the two scenarios, where the screen-based application may usually involve a task that is more intense and requires a higher concentration, and the users do not want to be disrupted.

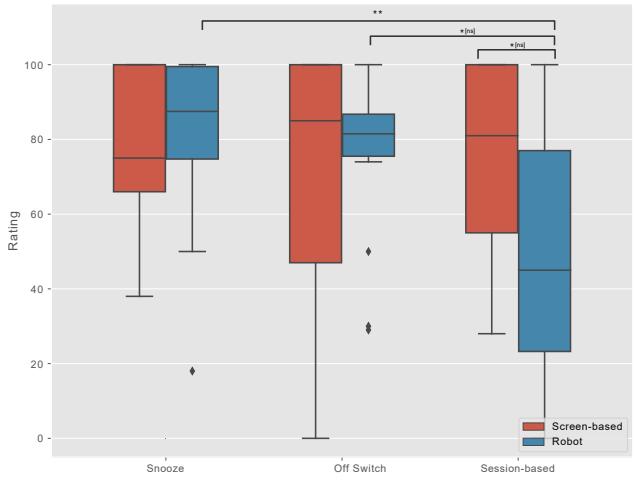

Figure 5: User Preferences: Features to Improve UX

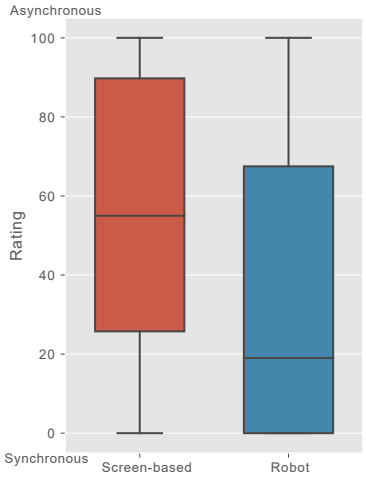

Figure 6: User Preferences: Feedback Timing

For example, P38 mentioned that they "*"would not want to be interrupted or distracted by feedback given when I'm in the middle of whatever I'm doing,*" and P7 asked "*Would it be able to tell when we are too busy to provide feedback at a particular moment?*" On the other hand, the robotic interaction might be imply a more casual interaction setting where the user and the robot is just having small talks, during which the disruption factor is less of a concern. Further investigations are required to better understand the underlying mechanisms. Regardless of deployment scenarios, participants also expressed reservations for the asynchronous approach, mostly due to concerns about accurate recall. For example, P9 mentioned that "*It may be hard to remember the emotion that was being felt if the question is asked later.*"

4.5 Additional User Feedback

Our questionnaire also collected qualitative data. While the questionnaire did not focus on open ended questions, users were given the opportunity to provide open feedback in several sections of the questionnaire as stated earlier. As these sections were optional not all users opted to leave feedback. We applied the basics of a thematic analysis.

Over the course of our questionnaire, we had an average response rate of 51% for the open response questions. Applying the basics of thematic analysis we determined a series of codes using the open coding paradigm. From these codes we were able to several themes that users were concerned about that are indirectly related to our research questions but are relevant given the sensitive nature of this topic. As these questions were not the focus of the study and the small response size we applied the principles of thematic analysis at a high level. We identified two major themes that were recurrent regardless of age, gender, or education level. These themes are privacy of data collected from participants and the comfort of participants with an AI device that can detect emotions. Of our participants 20% were concerned with data privacy and 31% indicated some kind of discomfort with the systems.

4.5.1 User Feedback Prompting. We found that FER system's deployment contexts can greatly influence the style of feedback prompting preferred by the users. Users preferred the textual prompting style in web or screen based applications, whereas users preferred the audio prompting style in robotic applications. We did not observe any difference in preference for the audio and textual prompting. These suggest that a certain user feedback elicitation modality is preferred by the users, and that it can depend on the application context. As for the user feedback elicitation frequencies, we found that users are willing to provide feedback to the system at a high frequency in the initial stage as users are beginning to get used to the system and as the system is learning more about the user's facial expressions. After a few sessions after they have gotten familiar with using the system, they prefer to be triggered less for user feedback. We also found that users generally find it very useful to have features like a manual switch that can disable feedback elicitation or a feature asking for a permission to elicit user feedback at the beginning of a session. Altogether, our findings suggest that developers of personalized FER model should be particularly mindful of the deployment context in eliciting user feedback. They should also be attentive to other factors such as the user's familiarity with the system or user's age range to meet user preferences in gathering feedback building a personalized FER model.

4.6 User Concerns

In our analysis section we also highlighted two main themes that concerned participants: privacy and level of comfort.

4.6.1 Privacy. The first theme concerns the privacy of participant data and the possible uses or misuses of such data. Participants expressed concern with government agencies having or being able to obtain access to this data or just general misuse of the sensitive training data. One participant expressed concern about how their data could be used for targeting advertising purposes:

> I would be concerned about potential uses by advertisers or other businesses. For example, would facial expressions be used by advertisers to evaluate reactions to products or online shopping?

4.6.2 Comfort. The second theme deals with the comfort of users having an AI system in their home. Participants expressed concern with feedback systems being disruptive to their everyday lives such as interrupting at inopportune times. Other participants expressed feelings of uneasiness with an artificial intelligence system analyzing them to determine the participants emotions:

> My first thought is that it creeps me out to think of AI attempting to understand people's emotions. I would not want it to be installed on any of my or my children's screens.

5 LIMITATIONS

There are several limitations to our current study. First of all, while we were able to recruit 38 participants in a short amount of time, all participants were recruited through friends and family, which may not be the most representative sample. However, given the diverse background of our team members (in terms of programs, degrees, and cultural background), we believe our respective network combined still lends itself to produce results that can be reasonably well generalized among the fairly well educated population aged between 20 and early 30s. Another limitation that we are aware of is the reliability of users self-reported preferences. A possible future work to address this concern is to develop systems to actually carry out these thought experiments, and more realistically evaluate users and their preferences in those settings. Finally, some ordering effects could be present as well. In hind sight, we would have randomized the order of the options within each group of questions. While we do not believe there is significant systematic bias as a result of such lack of randomization in the order of options, this could be further evaluated in a future replication that incorporates such measures.

6 CONCLUSION

In this project, we set out to understand the scope of considerations among users when it comes to building Personalized Facial Emotion Recognition systems. Specifically, we studied user preferences in terms of providing feedback to the system's emotion recognition performance. To that end, we designed and developed an online questionnaire grounded in both the field of human-robot interaction, and the field of human-AI interaction. Through friends and family, we recruited 38 participants, who were randomly assigned to one of the two hypothetical deployment scenarios: FER embedded within a robot, and as part of a screen-based application such as a conferencing app. Analyzing the responses, among other insights, we found that users generally prefer interacting verbally with robots, and textually with screen-based apps. We also summarized and presented common themes among user concerns, including privacy concerns, the ease of use, as well as a sense of uncomfort regarding FER systems at large. Based on these insights, we listed design implications, including the importance to consider the deployment context of the FER system when designing the primary mode of communication, as well as designing a system flexible enough to support user customization in several aspects that displayed high user preference variability. We urge future researchers to implement our findings to gather user feedback and investigate whether incorporating user feedback into a transfer learning mechanism can create a personalized FER system.

REFERENCES

[1] MAH Akhand, Shuvendu Roy, Nazmul Siddique, Md Abdus Samad Kamal, and Tetsuya Shimamura. 2021. Facial emotion recognition using transfer learning in the deep CNN. *Electronics* 10, 9 (2021), 1036.

[2] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[3] Kate Candon, Helen Zhou, Sarah Gillet, and Marynel Vázquez. 2023. Verbally Soliciting Human Feedback in Continuous Human-Robot Collaboration: Effects of the Framing and Timing of Reminders. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 290–300.

[4] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. 2022. Gender Stereotyping Impact in Facial Expression Recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 9–22.

[5] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[6] Zixiang Fei, Erfu Yang, David Day-Uei Li, Stephen Butler, Winifred Ijomah, Xia Li, and Huiyu Zhou. 2020. Deep convolution network based emotion analysis towards mental health care. *Neurocomputing* 388 (2020), 212–227.

[7] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2020. Explainable active learning (xal): An empirical study of how local explanations impact annotator experience. *arXiv preprint arXiv:2001.09219* (2020).

[8] M Rosario González-Rodríguez, M Carmen Díaz-Fernández, and Carmen Pacheco Gómez. 2020. Facial-expression recognition: An emergent approach to the measurement of tourist satisfaction through emotions. *Telematics and Informatics* 51 (2020), 101404.

[9] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. 2020. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 63–72.

[10] Ananya Ipsita, Hao Li, Runlin Duan, Yuanzhi Cao, Subramanian Chidambaram, Min Liu, and Karthik Ramani. 2021. VRFromX: from scanned reality to interactive virtual experience with human-in-the-loop. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[11] Yousif Khaireddin and Zhuofa Chen. 2021. Facial emotion recognition: State of the art performance on FER2013. *arXiv preprint arXiv:2105.03588* (2021).

[12] Eugenia Kim, De'Aira Bryant, Deepak Srikanth, and Ayanna Howard. 2021. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 638–644.

[13] Tianyi Li, Mihaela Vorvoreanu, Derek DeBellis, and Saleema Amershi. 2023. Assessing human-ai interaction early through factorial surveys: A study on the guidelines for human-ai interaction. *ACM Transactions on Computer-Human Interaction* 30, 5 (2023), 1–45.

[14] Mengqi Liao and S Shyam Sundar. 2021. How should AI systems talk to users when collecting their personal information? Effects of role framing and self-referencing on human-AI interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[15] Wafa Mellouk and Wahida Handouzi. 2020. Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science* 175 (2020), 689–694.

[16] Jaspar Pahl, Ines Rieger, Anna Möller, Thomas Wittenberg, and Ute Schmid. 2022. Female, white, 27? bias evaluation on data and algorithms for affect recognition in faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 973–987.

[17] E Pranav, Suraj Kamal, C Satheesh Chandran, and MH Supriya. 2020. Facial emotion recognition using deep convolutional neural network. In *2020 6th International conference on advanced computing and communication Systems (ICACCS)*. IEEE, 317–320.

[18] Burr Settles. 2011. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*. JMLR Workshop and Conference Proceedings, 1–18.

[19] Manali Shaha and Meenakshi Pawar. 2018. Transfer learning for image classification. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE, 656–660.

[20] Simone Stumpf, Erin Sullivan, Erin Fitzhenry, Ian Oberst, Weng-Keen Wong, and Margaret Burnett. 2008. Integrating rich user feedback into intelligent user interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*. 50–59.

[21] Güray Tonguç and Betul Ozaydın Ozkara. 2020. Automatic recognition of student emotions from facial expressions during a lecture. *Computers & Education* 148 (2020), 103797.
[22] Niels Van Berkel, Mikael B Skov, and Jesper Kjeldskov. 2021. Human-AI interaction: intermittent, continuous, and proactive. *Interactions* 28, 6 (2021), 67–71.
[23] Kees Van den Bos, Riel Vermunt, and Henk AM Wilke. 1996. The consistency rule and the voice effect: The influence of expectations on procedural fairness judgements and performance. *European Journal of Social Psychology* 26, 3 (1996), 411–428.
[24] Qiping Zhang, Austin Narcomey, Kate Candon, and Marynel Vázquez. 2023. Self-Annotation Methods for Aligning Implicit and Explicit Human Feedback in Human-Robot Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 398–407.