Detecting player emotions in serious games using AI-based recognition models

Cristina Alonso-Fernández
Software Engineering and Artificial Intelligence Departament
Complutense University of Madrid
Madrid, Spain
calonsofernandez@ucm.es

Antonio Calvo-Morata
Software Engineering and Artificial Intelligence Departament
Complutense University of Madrid
Madrid, Spain
acmorata@ucm.es

Baltasar Fernández-Manjón
Software Engineering and Artificial Intelligence Departament
Complutense University of Madrid
Madrid, Spain
balta@fdi.ucm.es

***Abstract*—The common evaluation of serious games is solely based on pre-post questionnaires and analytics collecting player interactions. However, other data sources like bio signals or players' emotions can help us to better understand and evaluate players' behavior. In this paper, we explore the use of AI-based facial emotion recognition (FER) models to detect player emotions while they are playing serious games. We initially evaluated the validity of this approach with a basic proof-of-concept with four users watching videos that clearly were prone to cause one emotion and comparing model outputs with manual annotations. After checking the validity of the FER models used, we conducted a pilot study with fifteen participants who played a serious game about sexism in the workplace. Both players' faces and screen gameplays were captured. Participants also completed a post-game questionnaire to self-report their feelings during the game. Results show a moderate relationship between detected and reported emotions, while also point out some limitations of the FER models. These early results suggest the feasibility of using FER models to complement the evaluation of serious games.**

Keywords—serious games, facial emotion recognition, learning analytics, game analytics.

I. Introduction

Serious games are a powerful tool to educate, train, and raise awareness about social issues [1]. Their evaluation is commonly performed based on pre-post questionnaires, which measure the impact of the game comparing metrics before and after playing it [2]. In addition to that, Game Learning Analytics (GLA) can also be used to capture players' interactions within the game and analyze them for different purposes, including players' assessment [3]. However, these evaluation methods do not consider the emotional experience of players during the gameplay [4]. These emotions can be related to engagement, learning, and attitude change. This may be particularly relevant in serious games that address social issues, in which players' emotions can be truly relevant to evaluating how much the game conveys the message that it intends to [5].

Facial emotion recognition (FER) techniques provide new opportunities to detect players' emotions while they are interacting with different environments, for instance learning tools [6]. These techniques have been used in different fields including the field of videogames [7], [8], and more specifically educational games [9].

However, several issues exist, such as technical constraints, complexity of the analysis, and individual differences in emotional expression, which make their application costly and sometimes ineffective.

This study proposes an AI-based approach to detect players' emotions while playing serious games. The approach combines the analysis of facial expressions captured via webcam, while also recording the gameplay on participants' computer screens. Our goal is to evaluate the feasibility and utility of using FER models to detect players' emotions while playing serious games where players play a leading role in a situation that may be uncomfortable. To evaluate it, we focus on a serious game designed to raise awareness about sexism in the workplace. We conducted this study in two steps: first, we conducted an initial validation of the models with short videos, followed by a proof-of-concept study using the serious game with 15 participants. Detected emotions were compared with the feelings that participants self-report in a post-game questionnaire.

The rest of this paper is structured as follows: Section 2 reviews related work on FER models and serious games evaluations; the methodology followed in our study is described in Section 3; its results are presented and consequently discussed in Section 4. Finally, in Section 5 we summarize the main conclusions, limitations, and lines of future work.

II. Related Work

Emotions play an essential role in the process of learning [10]. They influence not only learners' levels of engagement but also the depth and quality of their learning outcomes. Emotional states can facilitate or limit cognitive processes such as attention, memory, and motivation, all of which are fundamental to effective knowledge acquisition. This emotional dimension is also significant in digital learning environments, including serious games. In such contexts, games are often designed not merely for entertainment but to promote awareness, foster reflection, or encourage behavioral change regarding specific issues. Consequently, the emotions that players have during gameplay become an essential component in assessing games' impact [5].

Traditional evaluation approaches (such as pre- and post-questionnaires, interviews, or the analysis of interaction logs and game learning analytics) are limited when it comes to capturing emotional responses [4], [11]. These methods typically fail to assess emotions or only assess them after the gameplay is completed, providing only a partial understanding of players' in-game experiences. In particular, it is difficult to know at which point or scene in the game the players felt the emotions they described, resulting in only the predominant emotion during the session. As a result, critical insights into how players emotionally engage with key narrative moments, decision points, or learning challenges within the game are often lost. A more comprehensive evaluation framework that incorporates real-time emotional data could therefore provide a richer understanding of player experience, contributing to the design of more effective learning games.

Facial Emotion Recognition (FER) techniques, which integrate advances in artificial intelligence and computer vision, can detect and analyze users' emotional states in real time [6]. By interpreting facial expressions through automated image processing and machine learning algorithms, FER systems can provide objective and continuous measures of emotional responses. Due to these possibilities, FER techniques have been increasingly used across different domains, including education, human–computer interaction (HCI), and market research [12]. In these fields, FER techniques can contribute to understanding users' engagement and adjust the systems accordingly.

Despite this increasing application of FER across multiple domains, its application in the domain of serious games remains comparatively underexplored. Previous research in educational technologies has primarily used FER techniques to recognize learners' emotions in controlled environments or to enable adaptive responses in intelligent tutoring systems [13]. Such contexts, while informative, typically focus on structured learning tasks and may not capture the complex, spontaneous emotional dynamics elicited by interactive and narrative-driven game environments.

Recent works have started to investigate the potential of FER techniques to enhance the evaluation and adaptivity of serious games. These studies have demonstrated that integrating emotion recognition can enrich the understanding of user experience and offer objective, real-time feedback beyond traditional self-report instruments [14], [15]. For example, FER data can be used to detect moments of frustration, enjoyment, or surprise during gameplay, allowing researchers and designers to map emotional responses to specific in-game events or mechanics. This approach not only provides a more precise evaluation of player engagement but also supports the development of adaptive systems capable of responding dynamically to players' affective states.

This integration may be particularly relevant for serious games addressing complex social or emotional issues (for instance, serious games that aim to increase players' empathy, change prejudices, or address ethical aspects) where emotional impact constitutes a central dimension of the intended learning outcome. In contrast, educational games focused solely on learning content or procedures, may be more limited in terms of players' emotional responses.

For those reasons, we aim to evaluate the feasibility of using FER techniques to capture players' emotions while playing serious games particularly focusing on addressing social issues. Eventually, this could contribute to improving the evaluation of such serious games, providing an aspect key and not considered in traditional evaluation techniques (e.g., pre-post questionnaires or game analytics).

III. Methodology

The goal of this study is to evaluate the feasibility of using FER techniques to improve the validation of serious games. Building upon the identified research gap in the real-time assessment of emotions within serious games, this study was structured in two phases. The first phase aimed to conduct a preliminary validation of two FER models to examine their accuracy and feasibility in analyzing short video samples. The second phase employed the best-performing model in a proof-of-concept study using a serious game designed to raise awareness about sexism in the workplace.

The study was guided by the following research questions:

RQ1. To what extent existing facial emotion recognition (FER) models reliably detect emotional expressions in short video segments?

RQ2. Which model demonstrates the highest accuracy and stability detecting emotional expressions in short video segments?

RQ3. To what extent existing facial emotion recognition (FER) models reliably detect emotional expressions in a serious game context to capture players' emotions during gameplay?

RQ4. What insights into players' emotional engagement and experiential responses can FER data provide that complement or go beyond traditional self-report measures?

And consequently, the study was conducted in the corresponding two phases:

1. The initial preliminary validation of two different FER models, to evaluate the feasibility of the approach with short videos. It aims to respond to RQ1 and RQ2.
2. With the best-performing model, a proof-of-concept with a serious game to raise awareness about sexism in the workplace. It aims to respond to RQ3 and RQ4.

We present the methodology used in each of these phases in detail in the following subsections. First, we describe in detail in section 3.1 the initial validation with videos, as well as its results. The methodology applied in the proof-of-concept with a serious game is described in section 3.2, while its results are presented in Section 4.

A. Preliminary model validation with short videos

The goal of this initial phase is to verify the performance of the FER models and their capabilities to detect emotions. For that, we choose short videos that were prone to cause clear emotions in viewers, to check the performance of the FER models used to detect such emotions.

1) Validation setup

Four participants were included in this initial validation. The only inclusion criteria are that they were familiar with technology so that they could sit and watch videos on a PC.

The videos used in this validation were selected from YouTube, using the search bar and tags of the videos. The emotions represented in the videos were either tagged in them or stated in the description of the videos. In addition, they were checked by two of the study coordinators. 9 videos were selected that were aimed at causing the following emotions:

- Video 1: disgust. It shows unhygienic street food practices.
- Video 2: sadness or anger. It shows an aggressive bird trying to attack a baby animal.
- Video 3: joy. It shows a performer telling a joke.
- Video 4: fear or disgust. It shows a creepy SpongeBob fan creation.
- Video 5: joy. It shows a performer telling a joke.
- Video 6: sadness. It shows a child mourning his mother.

- Video 7: surprise or fear. It shows a cat jump-scare moment.
- Video 8: disgust. It shows a series of intentionally frustrating clips.
- Video 9: fear or joy. It shows a doll with a knife stopped by a funny kick. Therefore, it could cause either fear of the doll or laugh/joy for the ending.

Notice that in four of the videos two main emotions were represented, so we could test if the models were able to detect one or both, in case that participants showed them.

In addition to the videos, participants completed a questionnaire after watching them. In the questionnaire they were asked to select which emotion they felt while watching each of the videos. They could select from the seven main emotions: joy, sadness, disgust, anger, fear, surprise, or neutral.

The sessions had the following four steps: introduction, set-up, actual test, and questionnaire. In the introduction, participants were briefly told the goal of the study, highlighting that they must act naturally. Then, we had a brief set-up in which they were asked to show some emotions (joy, sadness, anger, surprise) to evaluate if the model is working. After those checks, the actual study started. The selected set of videos were shown to participants, while their expressions were being recorded with the PC webcam. The FER models were recording the emotions and storing them in a CSV file. After watching all videos, participants completed the post-test stating the emotion that they felt in each of them.

To verify the performance of the models, the study coordinators manually checked the recording of participants' faces to verify if the model was correctly detecting the changes in emotions, or if noise was changing the models' outputs.

For this initial validation, two different models were evaluated. 2 participants were evaluated with a FER model from Keras [16], while the other 2 used a PyTorch model [17]. The models store the output in a CSV file that contains: all 7 emotions and their probability, as well as the timestamp of the prediction. In order not to overload the output file, the following configurations were made: first, emotions were recorded only if they were sustained for several frames, and they were only stored in the CSV file if there has been a change from the previous emotion. These configurations reduce the amount of data to be recorded and simplify the application of FER models in actual gameplays.

2) Validation outcomes

Regarding the specific emotions detected, Table I and Table II present the summary of the emotions detected by the Keras and PyTorch models, respectively. In particular, for each video, we state the emotion expected in the video, the emotion participants stated in questionnaire, and the emotions detected by the model. In the final column, we evaluate whether the model correctly detects the expected emotion ("OK") or not ("FAIL") and provide additional information of the possible reason behind the failure of the model. These comments are based on the analysis of the participants' faces and the similarities between the facial characteristics of some emotions that may lead to models incorrectly recognizing them.

As seen in the tables, both models display similar performance in terms of emotions correctly detected. They also have similar limitations in emotion recognition, as both models failed to detect disgust and surprise. In addition, we observe that the "neural" emotion is commonly recognized and it provides no information, so we could discard it.

During the validation, we also noticed that participants' facial characteristics (e.g., facial hair) may influence the models' outcomes, as we detected more noise (i.e., less stability) in the models when evaluated with participants with beards.

Therefore, in response to **RQ1 (***To what extent existing facial emotion recognition (FER) models reliably detect emotional expressions in short video segments?*), results show that existing FER models have some limitations as they seem to most easily detect some specific emotions (joy, fear, sadness, anger) than others (disgust, surprise). They also detect the emotion "neutral" by default and need clear changes in facial expressions to detect other emotions than that by-default. The players' characteristics (e.g., facial hair, facial expressiveness) may also affect the results and limit the efficacy of such FER models.

In response to **RQ2** (*Which model demonstrates the highest accuracy and stability detecting emotional expressions in short video segments?)*, results in Table I and Table II show similar performance of both models. Both succeed in identifying the same emotions and fail to detect disgust and surprise.

Regarding their technical capabilities, Keras provides a high-level API that is easier to code, read, and execute across different frameworks, making it suitable for fast and intuitive model development. In contrast, PyTorch offers a low-level API that allows greater flexibility but requires more effort to reuse and maintain code. While TensorFlow/Keras is often preferred in industry for building scalable, production-ready models, PyTorch is widely used in research due to its adaptability and support for experimentation. Although it requires more CPU resources, PyTorch models offer superior speed and accuracy.

Combining all these reasons, we consider that the PyTorch model was more dependable, and we decided to choose it for the next proof-of-concept with a serious game.

B. Proof-of-concept with a serious game

After verifying the feasibility of the tested FER models to detect emotions, we conducted a proof-of-concept with an actual serious game. For that, we selected a game that aims to raise awareness about a social issue (gender discrimination in the workplace) and evaluated it with participants while their emotions were recorded.

1) Participants

Fifteen participants (14 male, 1 female) with basic technological skills were selected for the study. They were going to play a serious game so basic knowledge of computers was required. No other inclusion criteria were applied. Participants were informed about the data that was to be collected during the experiment (recording of their faces, screen capture of their gameplays) and gave their explicit consent to participate.

2) Materials

The materials used in the proof-of-concept consists of a serious game and a post-game questionnaire.

The serious game used in this proof-of-concept was called "La Entrevista" (Spanish for "The job interview") [18]. It is a point-and-click narrative game with basic mechanics (selections in multiple-choice options, conversations with NPCs). The goal of the game is to raise awareness about sexism in the workplace. For that, the story places players in first-person as someone going to conduct a job interview. They must go through several areas of the company (reception, waiting room, cafeteria, the room where the actual job interview occurs) where they meet different workers of the company (that are NPCs in the game), who they can interact with (see Fig. 1). In those conversations, they are exposed to different sexism situations, to which they can react.

Fig. 1. Screenshots of the game "La Entrevista" used to evaluate the models.

TABLE I. SUMMARY OF EMOTIONS IN EACH GAME SCENE DETECTED BY THE KERAS FER MODEL

Video	Expected emotion	Emotion stated in questionnaire	Detected emotions (Keras model)	Result	Comments
V1	Disgust	Disgust	Neutral, sadness, joy, fear, anger	FAIL	Possibly, it mixes anger and disgust
V2	Sadness or anger	Neutral	Neutral, sadness, joy, fear, anger	OK	Sadness and anger detected
V3	Joy	Joy	Neutral, sadness, joy, fear, anger	OK	Joy detected
V4	Fear or disgust	Surprise	Neutral, sadness, fear	OK	Fear detected
V5	Joy	Joy	Neutral, sadness, joy, fear	OK	Joy detected
V6	Sadness	Neutral	Neutral, sadness	OK	Sadness detected
V7	Surprise or fear	Neutral	Neutral, sadness	FAIL	No surprise or fear detected
V8	Anger	Anger	Neutral, sadness, joy, fear	FAIL	No anger detected
V9	Anger or joy	Joy	Neutral, sadness, joy, fear	OK	Joy detected

TABLE II. SUMMARY OF EMOTIONS IN EACH GAME SCENE DETECTED BY THE PYTORCH FER MODEL

Video	Expected emotion	Emotion stated in questionnaire	Detected emotions (PyTorch model)	Result	Comments
V1	Disgust	Disgust	Neutral, sadness, joy, fear	FAIL	Possibly, it mixes sadness and disgust
V2	Sadness or anger	Neutral	Neutral, sadness, joy, anger	OK	Sadness and anger detected
V3	Joy	Joy	Neutral, sadness, joy, anger	OK	Joy detected
V4	Fear or disgust	Surprise	Neutral, joy, anger	FAIL	Possibly, it mixes fear/disgust with anger
V5	Joy	Joy	Neutral, sadness, joy, anger	OK	Joy detected
V6	Sadness	Neutral	Neutral, joy, anger	FAIL	No sadness detected
V7	Surprise or fear	Neutral	Neutral, joy, anger	FAIL	Possibly, it mixes fear with anger
V8	Anger	Anger	Neutral, joy, anger	OK	Anger detected
V9	Anger or joy	Joy	Neutral, sadness, joy	OK	Joy detected

There are nine scenes in the game that present typical situations of sexism in the workplace. These scenes are:

- Game scene 1: at the reception desk, the receptionist thinks that the player is going to apply to a specific position in the company based solely on their appearance.
- Game scene 2: in the waiting room, a man discusses that some companies are starting to prefer hiring women or members of minority groups and express his disagreement with that.
- Game scene 3: in the waiting room, a man comments that he leaves all the household chores to his girlfriend because he does not know how to do them.
- Game scene 4: in the cafeteria, two men comment the lack of women in engineering and analysis positions, implying that they are incapable of having the same level of knowledge in that field.
- Game scene 5: in the cafeteria, in a conversation between two women, when a man appears and expresses his opinion, which is given more weight than the opinion of the woman.
- Game scene 6: right before the job interview starts in the interview room, the male interviewer insists that his female colleague should take notes during the interview, under the excuse that her handwriting is neater.
- Game scene 7: during the job interview, the male interviewer asks the candidate whether she has a partner.
- Game scene 8: during the job interview, the male interviewer comments that the candidate has a pleasant voice, which may be related to a weak attitude and low ability to manage pressure.
- Game scene 9: during the interview, the male interviewer assumes that the candidate must take care of the housework.

There is also a final key game scene in which the protagonist's characteristics are revealed to players. In that scene, the protagonist looks in a mirror, and players find out that they have been playing as a Black woman the whole time.

At the end of the game, the nine key game scenes are revised and the sexism situations they depict are described to players so they can review those situations and evaluate whether they were able to detect those situations or not.

During the design process of the game, those sexism situations were selected from a comparison and combination of both studies and reports of sexism in the workplace, and personal interview with female engineers. The game is expected to be completed in around 15 minutes [19].

Participants completed a post-game questionnaire after playing. The questionnaire explicitly asked players to report which emotion they felt the most during the game, and in each of the 9 key moments of the game. In these 10 questions, they could choose from the main emotions based on literature and the ones detected in the FER models: joy, sadness, disgust, anger, fear, surprise, or neutral. In addition, they were asked: whether they thought the character they were playing was a Black woman; which situations in the game were most surprising; and if they felt identified with any of the situations displayed in the game.

3) Session Design

The sessions had 3 steps: introduction, gameplay, and questionnaire. In the introduction, the study coordinators briefly explained the goal of the study, and that participants were about to play a serious game. Then, they moved to the actual gameplay where they started the game, and both the screen and participants' faces were recorded. After the game finished, participants completed the post-game questionnaire.

The setup for data collection was the following: we used the computer's webcam to record participants' faces, while the OBS Studio software was used to record the computers' screen with the actual gameplay. As both screen and webcam recordings had the timestamp, both recordings could be directly synchronized.

After completing the gameplay, the recording of participants' faces was evaluated with the FER model, as explained in the subsequent section. Finally, we compared the emotions reported in the post-game questionnaire with the emotions detected by the FER model. As we have the recording of the gameplay (in the computer screen) as well as participants' faces, we could match the key moments in the game (those nine key game scenes with sexism situations) to the emotions detected in those instants by the FER model. This allowed the comparison with the reported emotions in the post-game questionnaire.

4) FER models.

The PyTorch model was selected due to its better performance in the preliminary validation, as described previously. The recording of participants' faces was processed after completion, when it was input into the FER model to analyze it. As a result, the FER model outputs a CSV file with the detected emotions, alongside the duration and the initial time in which they were captured. As the videos had a large amount of data (15 minutes with around 30 FPS), we chose to store only the relevant information with the following adjustments: we only collected emotions that were sustained for at least half a second (to avoid noise) or 15 frames. To further reduce the size of the data collected, we record only emotions other than "neutral" as it is the default one identified by the models, as seen in the preliminary model validation.

Fig. 2 shows an example of the lines output in the CSV file with those three values (columns): emotion detected, duration of the emotion, and initial time in which it was captured. For instance, in the first line we can see that emotion "fear" is detected, it has a duration of 0.867 seconds, and that emotion has been recognized by the model after 7.933 seconds of starting the gameplay. Emotions display in Fig. 2 include "fear", "joy" (renamed "happy") and "anger" (renamed "angry"), although "neutral" and "sadness" emotions were also detected.

To give an example of the amount of information generated, in the evaluation with one participant that lasted 17 minutes and 54 seconds, the CSV file resulting of analyzing that video recording with the FER model output a total of 32216 emotion predictions, most of which were "neutral".

```
fear   ,0 days 00:00:00.867000,00:00:07:933000
happy  ,0 days 00:00:00.900000,00:00:10:933000
fear   ,0 days 00:00:00.667000,00:00:12:266000
happy  ,0 days 00:00:00.500000,00:00:13:766000
happy  ,0 days 00:00:01.234000,00:00:47:666000
happy  ,0 days 00:00:01.433000,00:00:49:733000
happy  ,0 days 00:00:01.533000,00:01:09:933000
happy  ,0 days 00:00:01.433000,00:01:13:100000
fear   ,0 days 00:00:00.700000,00:01:23:100000
happy  ,0 days 00:00:02.100000,00:01:29:600000
fear   ,0 days 00:00:00.566000,00:02:11:000000
happy  ,0 days 00:00:01.534000,00:02:13:866000
happy  ,0 days 00:00:00.700000,00:02:39:566000
angry  ,0 days 00:00:00.833000,00:03:12:100000
```

Fig. 2. Example of a CSV file generated by FER model using the game "La Entrevista." It provides a summary of emotions, duration, and initial time.

In more detail, we analyzed the emotions in the key 9 moments of the game. For the emotions detected by the FER model, we could extract the emotion detected by the model in each of these moments. This could be done as the output CSV file of emotions recognized provides the time passed since starting the gameplay, and we have the recording of the actual gameplays (recorded from the PC screen), so we could calculate the time passed until each player reached each of the key 9 game scenes. With that, we have information about the emotion detected by the FER model for each player in each key moment of the game. In addition to that, in the post-game questionnaire, we asked participants to state which emotion they felt in each of those moments.

Questionnaires' results show that the emotion that players felt the most throughout the game were neutral (5 participants), disgust (5), fear (1), and surprise (2). More than

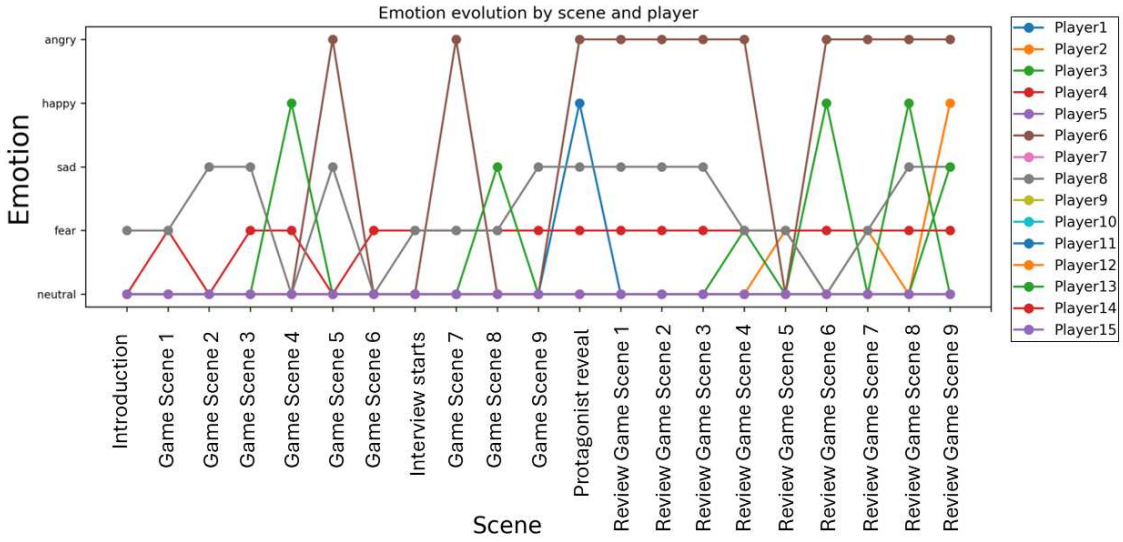

Fig. 3. Emotions detected by the FER model in each game scene.

IV. Results and Discussion

Emotions were detected and recorded by the model during gameplays, providing information in CSV files as the one displayed in Fig. 2. Those CSV files were then analyzed to see the change in emotions by players during the whole gameplay.

Fig. 3 displays those changes in emotions detected by the model. The recognized emotions appear in the y-axis (anger / joy / sadness / fear / neutral) aligned with all game scenes from beginning to end of the game (x-axis displays scenes chronologically). Game scenes include: the game introduction, the 9 game scenes providing relevant sexism situations (notice that in between those scenes 6 and 7 there is an additional scene recorded about the start of the job interview), the key scenes in which the protagonist's characteristics are revealed to players, and the final review of those key 9 game scenes. Each line in Fig. 3 represents one player. For instance, we can see that player tagged with "Player6" identifier displayed the "anger" emotion in game scenes 5 (conversation in the cafeteria) and 7 (inappropriate question in the job interview), and in many of the review scenes.

half of participants did not know and never would have imagined that they were playing as a Black woman, while the other half had doubts but were not sure at all. The most surprising situations in the game were personal questions in the job interview (related to Game Scene 7) and those related to Game Scene 2, 4 and 6.

The comparison of such emotions is displayed in Table 3. It presents the total emotions (for each of the possible emotions) detected by the models in each game scene (S1 to S9), compared to the emotions stated by participants in the questionnaire after playing. For instance, in the first key moment of the game (scene S1), the FER model detected the emotion "neutral" for twelve participants, while only seven participants stated to felt "neutral" while playing that scene in the post-game questionnaire. In one case the model and the questionnaire both detect the emotion "anger". Moreover, the FER model detected that two participants felt sadness and fear, and that was not stated in the post-game questionnaire. Conversely, two participants stated in the questionnaire to feel surprised and three to feel disgusted about this first game scene, and those emotions were not detected by the models.

TABLE III. NUMBER OF PARTICIPANTS THAT SHOW EACH EMOTION (COLUMN) IN EACH KEY GAME SCENE (ROW) AS DETECTED BY THE FER MODEL / STATED IN QUESTIONNAIRE

Game Scene	Neutral	Joy	Sadness	Disgust	Anger	Fear	Surprise
S1	12/7	0/0	1/0	0/3	1/1	1/0	0/2
S2	12/3	0/0	1/0	0/4	1/0	1/0	0/6
S3	12/3	0/0	1/0	0/7	1/0	1/0	0/3
S4	11/3	0/0	0/1	0/5	1/4	3/0	0/0
S5	12/7	0/0	0/2	0/0	0/0	3/0	0/4
S6	12/2	1/0	0/2	0/7	1/1	1/0	0/1
S7	11/1	0/0	0/0	0/4	1/2	3/0	0/6
S8	11/3	1/0	1/0	0/3	1/0	1/1	0/4
S9	10/2	1/0	2/0	0/5	1/2	1/0	0/4

In response to **RQ3** (*To what extent existing facial emotion recognition (FER) models reliably detect emotional expressions in a serious game context to capture players' emotions during gameplay?*), we again notice that the same limitations appear as in the case of short videos (RQ1). The models effectively detect some emotions (mainly: neutral, anger, or fear) and again fail to detect other emotions (disgust, and surprise).

In response to **RQ4** (*What insights into players' emotional engagement and experiential responses can FER data provide that complement or go beyond traditional self-report measures?),* results show that FER data can effectively detect some players' emotions and variations along gameplay. This is clear in Fig. 3, in which in a simple line chart we can see players' emotional changes alongside the gameplay. With this information, we could potentially relate game scenes that have a clearer impact on players' emotions (e.g., those scenes that cause players to feel anger or sadness), relate that to the sexism content of those particular game scenes, and relate that to players' responses in the post-game questionnaire. We could argue that the content depicted in those scenes may have had a bigger impact on the player – although more data needs to be included to be certain.

The insight given by players' emotional responses may also have the potential to improve the game design. For instance, we could analyze those scenes that have a lower-than-expected impact on players to revise their content and improve it. For instance, analyzing Fig. 3 we can see that Game Scene 2, about the discussion of companies' preference in hiring women or minorities, has only one player showing an emotional response (of sadness) to it.

V. CONCLUSIONS AND FUTURE WORK

The present study aims to evaluate the feasibility of using FER models to contribute to the evaluation of serious games. For that purpose, we conducted an initial validation of the models with short videos, to evaluate their performance in a context where the emotion that we expected each video to cause in participants was clear. That initial validation helped to evaluate the models, chose the best-performing one, and already detected issues that affect performance of FER models, and showed some of their limitations. Later, we conducted a proof-of-concept with a serious game to raise awareness about sexism in the workplace. The game was chosen to address a social issue that we expected to cause reactions in players, as they play a leading role in first person in a situation that may be uncomfortable.

Results, however, show several limitations. FER models failed to detect the emotions of "surprise" and "disgust" in participants. These could be caused by players' expressions not showing those emotions; however, these were commonly selected in participants' post-game questionnaires. The "neutral" emotion was easily and commonly detected by the FER models (as it is the default emotion). At least beforehand, this is not the most useful emotion to conduct an evaluation of a serious games – unless to say that the game did not cause a visible reaction in players. Nevertheless, we do notice that five participants stated in the post-game questionnaire that their main emotion during the gameplay was "neutral" so indeed the FER models may be correctly detecting those cases.

This exploratory study has several limitations: the limited number of participants, and the selection of only some videos and one serious game, limit the generalization of the results. To further evaluate the FER models, larger number of participants and more varied resources (e.g., different serious games) need to be evaluated.

Consequently, future lines of work are clear. We aim to evaluate the FER models in broader contexts (different serious games) and with a larger set of participants. The failure in detecting some particular emotions (surprise and disgust) also needs to be addressed, either testing the models in different scenarios to see if they are indeed able to detect those emotions in other cases or moving to different FER models that may be more complex and detect those emotions accurately.

Legal and ethical aspects also need to be addressed: the actual application of FER techniques in real contexts with users makes their application costly and risky, complying with all the data protection regulations (e.g., EU GDPR). To be able to universally apply these techniques to ensure that all privacy requirements are met, further work needs to be put into their simplification. For instance, evaluating players' emotions in real-time and only storing and analyzing the emotions (and not the actual videos of participants' face) avoiding capturing any personal data. This will ensure at least pseudo-anonymization

simplifying the actual application of these techniques in experiments and facilitating evaluations with users.

ACKNOWLEDGMENT

This work was partially funded by the Ministry of Education (PID2020-119620RB-I00; PID2023-149341OB-I00), and by the Telefonica-Complutense Honorary Chair on Digital Education and Serious Games.

REFERENCES

[1] D. R. Michael and S. L. Chen, Serious Games: Games That Educate, Train, and In-form. Muska & Lipman/Premier-Trade, 2005.

[2] A. Calderón and M. Ruiz, "A systematic literature review on serious games evaluation: An application to software project management," Comput Educ, vol. 87, pp. 396–422, Sep. 2015, doi: 10.1016/j.compedu.2015.07.011.

[3] C. Alonso-Fernández, A. Calvo-Morata, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón, "Applications of data science to game learning analytics data: A systematic literature review," Comput Educ, vol. 141, p. 103612, Nov. 2019, doi: 10.1016/j.compedu.2019.103612.K. Elissa, "Title of paper if known," unpublished.

[4] A. Hamrouni and F. Bendella, "Recognizing students emotions in game-based learning environment," International Journal of Information Technology, vol. 17, no. 6, pp. 3465–3475, Jul. 2025, doi: 10.1007/s41870-024-01802-4.

[5] L. Wei et al., "Amplifying Player Experience to Facilitate Prosocial Outcomes in a Narrative-Based Serious Game," Media Commun, vol. 13, Feb. 2025, doi: 10.17645/mac.8637.

[6] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar, "Understanding Deep Learning Techniques for Recognition of Human Emotions Using Facial Expressions: A Comprehensive Survey," IEEE Trans Instrum Meas, vol. 72, pp. 1–31, 2023, doi: 10.1109/TIM.2023.3243661.

[7] O. AlZoubi, B. AlMakhadmeh, M. Bani Yassein, and W. Mardini, "Detecting naturalistic expression of emotions using physiological signals while playing video games," J Ambient Intell Humaniz Comput, vol. 14, no. 2, pp. 1133–1146, Feb. 2023, doi: 10.1007/s12652-021-03367-7.

[8] Akbar, M. T., Ilmi, M. N., Rumayar, I. v., Moniaga, J., Chen, T.-K., & Chowanda, A. (2019). Enhancing Game Experience with Facial Expression Recognition as Dynamic Balancing. Procedia Computer Science, 157, 388–395. https://doi.org/10.1016/j.procs.2019.08.230

[9] Wiklund, M., Rudenmalm, W., Norberg, L., Westin, T., & Mozelius, P. (2015). Evaluating Educational Games Using Facial Expression Recognition Software: Measurement of Gaming Emotion. https://login.bucm.idm.oclc.org/login?url=https://www.proquest.com/conference-papers-proceedings/evaluating-educational-games-using-facial/docview/1728409740/se-2?accountid=14514

[10] C. M. Tyng, H. U. Amin, M. N. M. Saad, and A. S. Malik, "The Influences of Emotion on Learning and Memory," Front Psychol, vol. 8, Aug. 2017, doi: 10.3389/fpsyg.2017.01454.

[11] M. Freire, Á. Serrano-Laguna, B. M. Iglesias, I. Martínez-Ortiz, P. Moreno-Ger, and B. Fernández-Manjón, "Game Learning Analytics: Learning Analytics for Serious Games," in Learning, Design, and Technology, Cham: Springer International Publishing, 2016, pp. 1–29. doi: 10.1007/978-3-319-17727-4_21-1.

[12] [X. Zhao, J. Zhu, B. Luo, and Y. Gao, "Survey on Facial Expression Recognition: History, Applications, and Challenges," IEEE MultiMedia, vol. 28, no. 4, pp. 38–44, Oct. 2021, doi: 10.1109/MMUL.2021.3107862.

[13] A. V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying Emotions and Engagement in Online Learning Based on a Single Facial Expression Recognition Neural Network," IEEE Trans Affect Comput, vol. 13, no. 4, pp. 2132–2143, Oct. 2022, doi: 10.1109/TAFFC.2022.3188390.

[14] B. Mostefai, A. Balla, and P. Trigano, "A generic and efficient emotion-driven approach toward personalized assessment and adaptation in serious games," Cogn Syst Res, vol. 56, pp. 82–106, Aug. 2019, doi: 10.1016/j.cogsys.2019.03.006.

[15] S.-Y. Deng and K.-K. Fan, "Evaluation System for Game Playability Using Emotion Sensor Based on AI," Sensors and Materials, vol. 33, no. 9, p. 3379, Sep. 2021, doi: 10.18494/SAM.2021.3479.

[16] Serengil, S., & Özpınar, A. (2024). A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules. Bilişim Teknolojileri Dergisi, 17(2), 95–107. https://doi.org/10.17671/gazibtd.1399077

[17] Tr Pakov "Vision Transformer (ViT) for Facial Expression Recognition Model Card". HuggingFace, 2025 https://huggingface.co/trpakov/vit-face-expression

[18] A. Calvo-Morata and B. Fernández-Manjón, "Serious Games for Social Problems," Lecture Notes in Computer Science. Springer International Publishing, pp. 98–109, 2023. doi: 10.1007/978-3-031-33023-0_9.

[19] A. Calvo-Morata, C. Alonso-Fernández, B. Fernández-Manjón. (2025). Design of a Serious Game to Challenge Sexism. In: Sugimoto, M., Di Iorio, A., Figueroa, P., Yamanishi, R., Matsumura, K. (eds) Entertainment Computing – ICEC 2025. ICEC 2025. Lecture Notes in Computer Science, vol 16042. Springer, Cham. https://doi.org/10.1007/978-3-032-02555-5_4