Find the Bot!: Gamifying Facial Emotion Recognition for Both Human Training and Machine Learning Data Collection

Yeonsun Yang
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
diddustjs98@dgist.ac.kr

Ahyeon Shin
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
ahyeon@dgist.ac.kr

Nayoung Kim
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
skdud727@dgist.ac.kr

Huidam Woo
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
huidamwoo@dgist.ac.kr

John Joon Young Chung
Midjourney.Inc
San Francisco, California, United States
jchung@midjourney.com

Jean Y. Song
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
jeansong@dgist.ac.kr

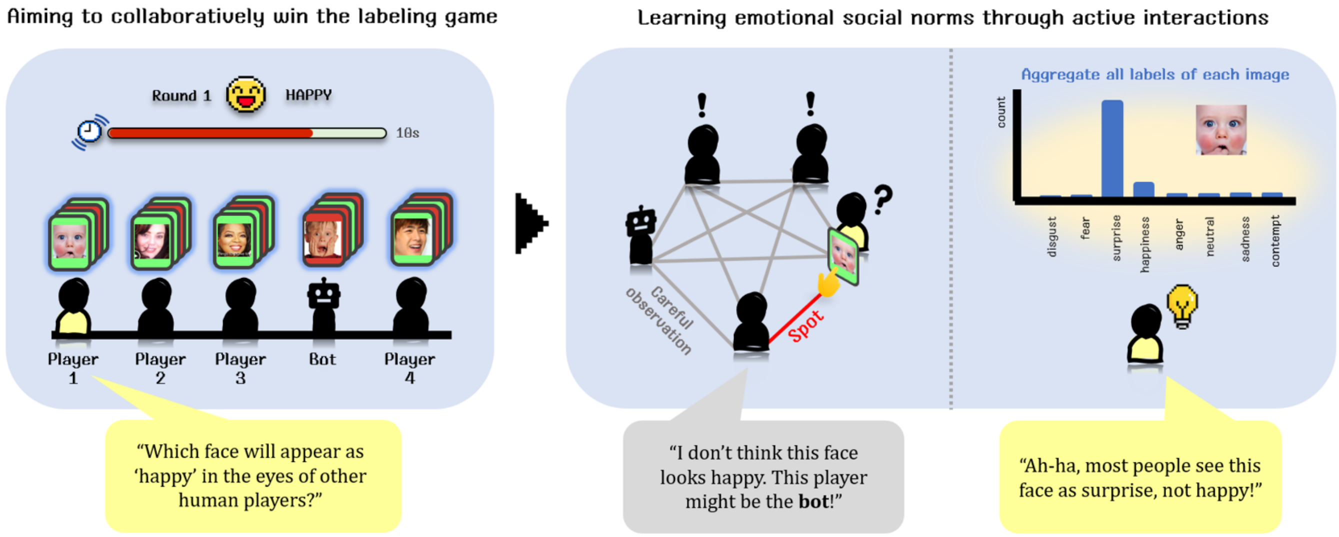

Aiming to collaboratively win the labeling game
Round 1
HAPPY
10s
Player 1
Player 2
Player 3
Bot
Player 4
"Which face will appear as 'happy' in the eyes of other human players?"
Learning emotional social norms through active interactions
Careful observation
Spot
"I don't think this face looks happy. This player might be the bot!"
Aggregate all labels of each image
count
disgust
fear
surprise
happiness
anger
neutral
sadness
contempt
"Ah-ha, most people see this face as surprise, not happy!"

Figure 1: Find the Bot! is an interactive and collaborative web-based game that enables players to naturally enhance their facial emotion recognition skills while also contributing rich, reliable emotion labels for training machine learning models—all while striving to win the game. The game design promotes observational learning and offers real-time personalized feedback from other players, facilitating the training of socially agreed-upon emotion labels, even for spontaneous and ambiguous facial expression images.

ABSTRACT

Facial emotion recognition (FER) constitutes an essential social skill for both humans and machines to interact with others. To this end, computer interfaces serve as valuable tools for training individuals to improve FER abilities, while also serving as tools for gathering labels to train FER machine learning datasets. However, existing tools have limitations on the scope and methods of training non-clinical populations and also on collecting labels for machines. In this study, we introduce Find the Bot!, an integrated game that effectively engages the general population to support not only human FER learning on spontaneous expressions but also the collection of reliable judgment-based labels. We incorporated design guidelines from gamification, education, and crowdsourcing literature to engage and motivate players. Our evaluation (N=59) shows that the game encourages players to learn emotional social norms on perceived facial expressions with a high agreement rate, facilitating effective FER learning and reliable label collection all while enjoying gameplay.

CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642880

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI); Human computer interaction (HCI)**.

KEYWORDS

gamification, facial emotion recognition (FER), computer interfaces for training, dataset collection

ACM Reference Format:
Yeonsun Yang, Ahyeon Shin, Nayoung Kim, Huidam Woo, John Joon Young Chung, and Jean Y. Song. 2024. Find the Bot!: Gamifying Facial Emotion Recognition for Both Human Training and Machine Learning Data Collection. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3613904.3642880

1 INTRODUCTION

The ability to accurately recognize the emotions of others by observing their facial expressions, known as facial emotion recognition (FER), plays a crucial role for both humans and machines, affecting the experience of not only human-human interactions but also human-machine interactions. In the context of human-human interaction, higher FER ability is associated with various psychosocial benefits, eventually improving academic and workplace performance [12, 31, 32, 62]. For machines interacting with people, it becomes possible to provide personalized services when equipped with FER technologies— catering to people's specific needs and preferences [11, 34, 55, 82].

With the advancement in software technology and the exponential increase in online interactions, various computer-based training tools that support the acquisition of FER abilities like Micro-Expressions Training Tool (METT) [23] or Emotion Trainer [67] have been introduced. These tools serve as a scalable and flexible alternative to conventional in-person training programs [20, 27, 54]. Most of existing tools are targeted to train a specific group of people, such as clinical populations with autism or Asperger syndrome, or police and security personnel who need special training on reading micro expressions. These tools typically use a sign-based approach where the facial expressions are divided and interpreted into small action units (e.g., cheek raiser, lip corner puller, or nose wrinkler) [24] that are manually coded by experts in advance. However, we argue that the sign-based approach is not appropriate for helping the general and non-clinical populations who would benefit from *naturally* learning diverse and nuanced facial expressions. We believe that a judgment-based approach [35, 49] may be a more practical training material for general people because it interprets facial expression based on how it is universally and heuristically perceived by a large common population, capturing emotional social norms shared among general people.

Meanwhile, collecting FER datasets to train artificial intelligence (AI) requires computer-based tools to create image-label pairs. Usually, paid annotators are recruited through crowdsourcing to work on a web-based interface where the interface presents facial expression images alongside labeling tools. Although machine learning researchers have recently been focusing on constructing large-scale datasets containing in-the-wild images that are labeled with a judgment-based approach [18, 29, 58, 59, 87], these datasets still exhibit limitations on reliability and robustness. This is primarily because such datasets are often annotated by a limited number of untrained annotators, causing issues related to personal biases and labeling errors. Instead, we suggest creating reliable FER datasets by involving a broader population that can provide socially agreed labels, which could yield greater advantages in training AI agents, particularly for those that have to interpret diverse people's emotions in real-world scenarios.

In this work, we pay attention to the fact that appropriate training materials are key requirements for both human FER training and machine learning data collection. To this end, we present *Find the Bot!* (Figure 1), a web-based game that engages a group of the general population to naturally train individuals with their FER ability while enjoying the game, and simultaneously obtain rich judgment-based annotations that can be used later to train other AI algorithms on in-the-wild facial images with socially agreed upon emotion labels. Our game is inspired by the globally popular game of Mafia (also known as "Werewolf"'), where players cooperate to find the 'Mafia' among themselves through active interactions such as observation, debating, and voting.[1] We hypothesize that this collaborative and immersive mainstream game can effectively address additional challenges identified in current training and data collection interfaces by seamlessly incorporating a wide range of suggestions and guidelines from gamification, education, and crowdsourcing into a single application. Specifically, we set the following research questions:

- RQ1 : Does Find the Bot! provide an engaging game experience to all players?
- RQ2 : Does Find the Bot! increase judgment-based FER scores for players who had low FER scores?
- RQ3 : Does Find the Bot! increase the social agreement of the collected labels on facial expression images?

To evaluate the feasibility and user experience of Find the Bot! and answer to the research questions, we conducted a user study with 59 participants, where we classified 22 of them as low FER group based on their pre-survey FER scores. We randomly divided the low FER group into learner group (N=11) who used Find the Bot! and control group (N=11) who did not, and compared the changes in their pre- and post-FER scores. We also evaluated the quality of the labels collected in the game. In addition, we qualitatively analyzed measurements of usability (SUS) and game experience (GEQ and custom questionnaires) in a post-survey. Our results suggest that Find the Bot! effectively helps train the non-clinical population with low FER abilities and helps collect reliable and socially agreed-upon labels through a well-motivated combination of game elements.

In sum, this paper makes the following contributions:

- We investigate and summarize the primary limitations in the design of both existing human FER training interfaces and FER machine learning dataset collection interfaces.
- We present the design and implementation of an interactive web-based game, Find the Bot!, that adopts findings from literature in gamification, education, and crowdsourcing to improve the performance of FER training on non-clinical populations and FER dataset collection for later AI training.

[1] We released Find the Bot! as an open-source repository for further research: https://github.com/diag-dgist/FindtheBot.

Challenges in Current Human FER Training Interfaces	
Limited to sign-based training [20, 27, 54]	Facial expression images taken in controlled environments with AUs are hard to capture diverse and ambiguous facial expressions in the real world.
Limited to self-administered training [23, 67]	Self-administered training on a computer is less effective than administered by a human instructor or in small groups.
Limited to partially-combined sessions [7]	A fraction of the necessary sessions (i.e., instruction, practice, and feedback) is less effective than a combination of all these sessions.
Tedious and repetitive sessions [5, 50, 64]	Simple task design demotivates learners from consistent and effective training.
Challenges in Current Machine Learning Data Collection Interfaces	
Large labeling error and bias [18, 29]	Limited number of annotators are prone to make biases and erroneous decisions, especially when they are untrained and crowdsourced.
Limited to single-choice format [13]	Single-choice interface design makes it difficult to annotate facial expressions that are complex and ambiguous, especially when lacking sufficient contextual information for annotation with a single label.
Limited to efficient estimation [13]	Crowdsourcing answer distributions requires a large number of answers to be collected to form a stable distribution, resulting in high expenses.
Tedious and repetitive sessions [48, 71, 72]	Simple task design demotivates annotators, resulting in careless and poor label quality.

Table 1: Challenges in current human FER training interfaces and machine learning data collection interfaces, which hinder the effective utilization of real-world spontaneous facial expression images that can teach various possible interpretations of the expressions.

- We report results from a controlled user study demonstrating that Find the Bot! facilitates consensus on emotional perceptions through active interactions, benefiting both learning of emotional social norms and the quality of collected labels.
- We identify the effectiveness of specific elements within our game and offer insights and recommendations for designing engaging and motivating games with a purpose.

2 BACKGROUND AND RELATED WORK

In this paper, we aim to design a web-based game that attracts people to enjoy, and as a byproduct, supports effective FER training and reliable label collection on spontaneous facial expression images. We review related work and their limitations in (1) FER training for humans and (2) FER labeling for machines, which is summarized in Table 1. Then, we investigate the landscape and design principles for web-based games, which lead us to design a game that combats the limitations of both (1) and (2) within a single game design. Finally, we examine the elements for an effective learning process in general, aiming to enhance FER learning effectiveness through user interactions within our designed game.

2.1 FER Training for Humans

FER has been widely studied for decades in psychology. Extensive research has shown that FER abilities play a crucial role in our daily social lives. For example, higher FER is associated with various psychosocial benefits, including better relationship quality [21, 30, 39, 80], social functioning [25, 36, 37, 45], and eventually improving academic and workplace performance [12, 31, 32, 62]. While psychologists have made significant efforts to improve individual FER by offering diverse training methods, ranging from in-person training with one instructor or in small groups [20, 27, 54] to self-administered computer-based training [61, 66, 67, 83, 85], these methods are designed to train specific groups of people who have clinically significant deficits in FER abilities, or who need professional training to be more sensitive to recognizing micro expressive emotions. We note that the goal of our work in this paper is different from this previous work because we aim to build an interactive and fun game that helps train FER for the general and non-clinical population, which can positively impact their workplace performance and everyday social interactions.

Most in-person training has limitations in terms of scalability and flexibility. For example, learners must schedule appointments with trainers, visit periodically, and pay for the treatment. Moreover, human trainers can only physically accommodate a limited number of learners, and the effectiveness of the training heavily depends on their skill and experience. Thus, FER training is increasingly offered through computer interfaces, such as Micro-Expressions Training Tool (METT) [23], Emotion Trainer [67], and others [5, 50, 64]. To reduce the dependence on human instructors, such tools are designed to provide fully automated instruction, practice, and feedback sessions. For example, learners are taught by written instructions for each emotion expression. Then, with the instructions in mind, they practice making emotional assumptions by viewing facial expression images on the screen, clicking on a correct emotion label button, and receiving feedback such as 'well done' or 'try again'.

However, the standardized and posed facial expression images used in these computer-based tools remain a limitation when training the general population with diverse, nuanced, and spontaneous facial expressions. For example, Barrett et al. [3] argued that these sign-based images with manually coded action units (AUs) are limited to prototypical facial expressions rather than capturing and showing the ambiguous and complex aspects of spontaneous facial expressions. In addition, facial expressions posed by the actors are mostly expressed as exaggerated and only applicable to limited social contexts (e.g., only shows prototypical expressions of American people where the actors were hired). It also requires a significant amount of human effort to record the posed images and manually code the AUs by experts.

In contrast, a judgment-based approach [35, 49] considers the common expression perception by a large population as the gold standard emotions, instead of the professionally coded AUs of a few experts. It is suggested that reliable labels on spontaneous facial expression images could be better collected through the judgment-based approach [35]. We believe that the judgment-based approach is a more suitable approach for training FER for the non-clinical populations because it captures the emotional social norms shared among general people and helps interpret facial expression based on how it is universally and heuristically perceived by a large common population.

2.2 FER Labeling for Machines Learning Datasets

With affective computing applications on the rise, the use of automatic FER machines has become commonplace in many contexts such as security [63], driver safety [40], healthcare systems [51, 73], and others [56, 86]. During the last decades, machine learning researchers have made promising progress in building datasets and developing AI models for FER [18, 29]. Due to a shift from model-centric to data-centric AI [60], researchers also started to build diverse training datasets with in-the-wild images [58, 59, 87].

Despite these efforts, most existing datasets still do not capture the diverse interpretations of facial interpretations. That is, these existing FER datasets do not fully incorporate the complexity, ambiguity, and subjectivity found in spontaneous facial expressions. This is because the images are typically annotated by only one or two annotators assigned per image, who are given a single-choice interface to annotate an emotion label to an image (e.g., providing a radio button UI to choose a single emotion that represents the given facial image). This precludes the possibility of secondary or subtle emotions being labeled on facial images and excludes the images with ambiguous expressions from the final dataset even if they can be useful in training AI agents [13]. In this work, we present the design of an interface that displays an emotion keyword and then prompts users to provide binary labels on facial expression images (to indicate whether the emotion 'exists' or 'not exists' in the image). By aggregating these emotion-specific labels, we capture diverse valid interpretations for even ambiguous or complex facial expressions.

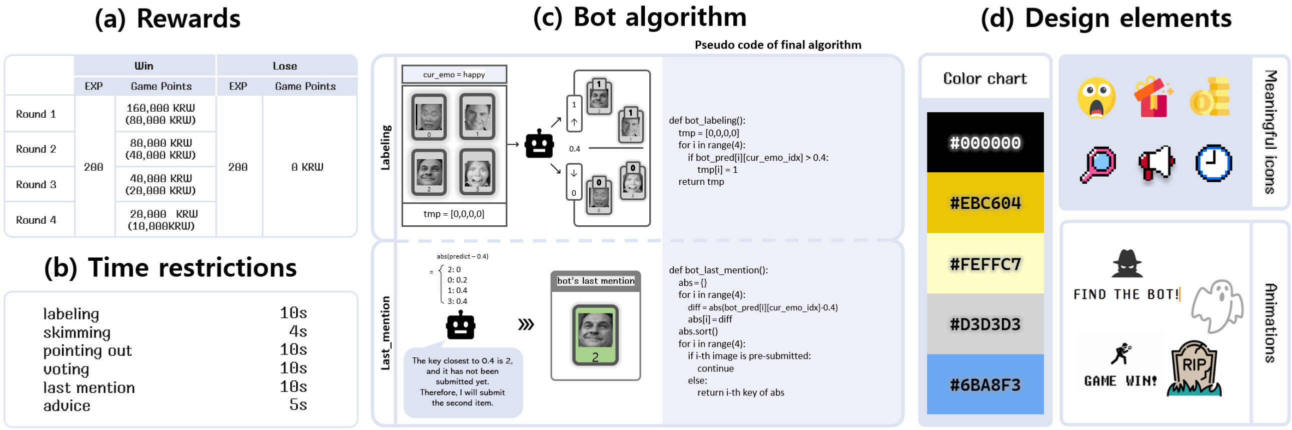
Quality control in label collection has long been a challenge for many researchers [48, 71, 72]. A typical approach to improve annotation quality is to collect multiple valid responses from a large number of annotators and aggregate them with answer distributions, particularly in domains where answers are ambiguous or subjective. Crowdsourcing researchers have found that a set of non-expert's aggregated annotations can outperform the quality of a single expert because annotators' diversity can help mitigate an individual's bias and subjectivity [69]. However, collecting annotations or labels that are socially agreed upon, especially for those that are complex and ambiguous, is a challenging task. It requires responses from a large population representative of society, thereby incurring high expenses in terms of human effort and cost [13].

One way to create an effect of recruiting a large group of annotators with just a few is to elicit richer responses from each worker by asking them to estimate the label distribution of a larger group and providing them with an interface to choose all plausible labels instead of forcing them to choose a single label (i.e., multi labeling and peer prediction from Bayesian Truth Serum) [13]. Inspired by previous work, we explore the design of a game that can engage players to naturally provide FER labels while trying to win the game. Our game design induces players to guess other players' perspectives to avoid being suspected as the hidden 'bot'.

2.3 Games with a Purpose

Games with a purpose can serve as an effective tool for engaging and motivating a large group of people for both peer learning and dataset collection. Gamification has been proposed in various domains and some with great success, such as protein folding with FoldIt [15], classification of galaxies with Galaxy Zoo [42], collecting common-sense knowledge with Verbosity [78], and others [77, 79]. To improve the effectiveness of FER training by encouraging learner motivation and attention, few studies have introduced gamified FER training tools: Let's face it! [74], Junior Detective Training Program [5], and MT-ALEX [50]. While such gameplay (e.g., shooting games or role-playing games) can be engaging, they are designed specifically for children with autism, Asperger's syndrome or for alexithymic individuals, making it challenging to generalize their effectiveness to adults and non-clinical populations. Building on this prior work, we aim to extend the benefits of games to a wider range of people.

Meanwhile, collecting labels for AI training during gaming can be an efficient approach to enrich the dataset. For example, ESP game [76] is a two-player online game that collects web image labels through player consensus on image descriptions. ASL Sea Battle [8] is a sign language game designed to collect ASL videos and labels while educating users. We are inspired by these existing practices



(a) Rewards

	Win		Lose	
	EXP	Game Points	EXP	Game Points
Round 1	200	160,000 KRW (80,000 KRW)	200	0 KRW
Round 2		80,000 KRW (40,000 KRW)		
Round 3		40,000 KRW (20,000 KRW)		
Round 4		20,000 KRW (10,000KRW)		

(b) Time restrictions

labeling	10s
skimming	4s
pointing out	10s
voting	10s
last mention	10s
advice	5s

```
def bot_labeling():
    tmp = [0,0,0,0]
    for i in range(4):
        if bot_pred[i][cur_emo_idx] > 0.4:
            tmp[i] = 1
    return tmp
```

```
def bot_last_mention():
    abs = {}
    for i in range(4):
        diff = abs(bot_pred[i][cur_emo_idx]-0.4)
        abs[i] = diff
    abs.sort()
    for i in range(4):
        if i-th image is pre-submitted:
            continue
        else:
            return i-th key of abs
```

Figure 2: Design elements and Bot's algorithm decided in the final design probe session. (a) The rewards are halved in each subsequent round. The amounts in parentheses indicate the halved rewards imposed as a penalty for players who are deactivated midway through the game. Experience points are consistently awarded at a rate of 200 EXP, regardless of whether players win or lose. Parameters such as (b) Time restrictions and (c) the threshold for the bot's prediction accuracy in labeling and last mention were decided based on feedback from the iterative design probes. (d) A color chart is used to provide a consistent look and feel, and friendly emojis are included to increase engagement.

of collecting high-quality labels during interactive and intensive gameplay. However, because the applicable range of existing game designs is limited to annotating only objective data with obvious ground truth, we propose a novel game design to collect reliable labels for subjective facial expression data by integrating advanced crowdsourcing techniques within a web-based game interface.

The right combination of game elements is a key requirement for engaging people, motivating action, promoting learning and solving problems [43]. Drawing from various gamification strategies discussed in literature [17, 19, 44, 70], we identified a comprehensive list of elements and incorporated them into the design of Find the Bot! through three rounds of design probing, which are detailed in Section 3.

2.4 Design Elements for Effective Learning

The successful integration of design elements is crucial in creating effective learning tools, especially in online learning environments. An early work on e-learning has explored six elements of effective design, which are Activity, Scenario, Feedback, Delivery, Context, and Influence [10]. Another work suggests that Presentation, Hypermediality, Application Proactivity, and User's Activity are the core dimensions for evaluating e-learning tools [2].

User activity element in e-learning involves interactive tasks or exercises that learners engage with to reinforce and apply their knowledge. These can include quizzes, simulations, discussions, and other interactive elements [68]. The *presentation and delivery element* (including hypermediality) focuses on how the content is presented to learners. This includes the overall design, layout, and multimedia elements such as text, images, and videos. Effective delivery ensures that the experience is engaging. The *scenario and context elements* suggest creating realistic contexts for learning by presenting learners with situations or challenges they might encounter in the workplace or in practical scenarios [14]. Learners navigate through these scenarios, making decisions and experiencing consequences. *Feedback and application proactivity elements* are crucial for learners to understand their performance [16]. In e-learning, feedback can be immediate, providing guidance on correct or incorrect answers, and motivating them to reflect on. Overall, when these design elements are combined thoughtfully, they contribute to a comprehensive and engaging learning environment.

In the domain of FER learning, previous research has shown that training for an individual's FER perception is more effective when administered by a human instructor or in small groups rather than being self-administered on a computer [7]. Unfortunately, current FER training tools [5, 23, 74] lack human interaction due to their exclusive focus on self-administration through fully-automated processes. Moreover, some tools provide only a fraction of the necessary sessions consisting of instruction, practice, and feedback, rather than combining them to create a comprehensive training experience [7]. This hinders the trainee from getting feedback from others or going through multiple effective practice sessions.

In this work, we aim to not only include all three necessary training sessions for FER learning (instruction, practice, and feedback sessions) into our proposed game design but also aim to incorporate the design elements for effective e-learning suggested by prior work. We came up with 14 detailed game elements grouped into 10 based on an overarching motivational strategy from previous studies [2, 10, 17, 19, 44, 70]. The element groups include a storyline, social pressure (both related to *scenario and context* element), challenge, competition (both related to *user activity* element), reminder, aesthetic (both related to *presentation and delivery* element), progression, reward, status, and punishment (all related to *feedback and* *application proactivity* element), which are summarized in Table 8 in Appendix A.

3 DESIGN PROBING

In this work, we design an integrated game that can solve two major challenges within a single application: (1) challenge in interface design for human FER training and (2) challenge in interface design for machine learning FER dataset labeling. More specifically, we set our first design goal as follows:

- DG1: Enable diverse layers of interactions between players so that they can learn socially agreed-upon interpretations of emotions through observational learning and real-time personalized feedback from others.

Through an iterative design process where we prototyped different interface mock-ups, we formulated a more specific set of design goals that helped resolve practical challenges we encountered during prototyping. For each prototyping iteration, we reflected feedback from three graduate students and six undergraduate students recruited in our institution (7 male, 2 female; age M=22.67 and SD=2.18). Each person participated to test a prototype for approximately an hour and a half providing verbal feedback in think aloud protocol. The added design goals are as follows:

- DG2: Minimize the difficulty and effort required for the labeling actions during the game so that the players can focus on the gameplay.
- DG3: Provide game rules and elements that are easy to learn so that not only those who desire to improve their FER skills but also ordinary players can effectively engage and enjoy winning the game.

The below subsection summarizes what we found through the iterative design probes and how we incorporated them into our final design of Find the Bot!.

3.1 Findings and Design Considerations

3.1.1 Using a storyline and sophisticated reward systems for integrating complex game elements: In the first design probe, we tailored ESP game [76] format, but the overly simplified design hindered participants from having a gameful experience. On the other hand, the addition of too many design elements specified for FER learning in our second design probe failed to motivate participants due to the complexity. From the literature survey, we found that an easy-to-understand scenario could alleviate the complexity, which led us to include a Mafia game storyline of finding a non-human bot to closely tie all necessary game elements without hindering game coherence.

We also found that providing sophisticated rewards system motivates users and helps maintain their engagement [44]. For more realistic and sophisticated rewards system, we used actual currency as points, which participants reported to provide stronger motivation toward gameplay. Additionally, points were designed to be rewarded differently based on various scenarios in our final design (as shown in Figure 2(a)).

3.1.2 Applying microtask workflow to lower task complexity: In our initial prototype, participants were required to click all plausible emotion labels for a single image within a limited time frame, which most participants found stressful and led to poor labeling quality. Consequently, we ended up breaking this complex task into smaller units of work (as done in microtask crowdsourcing), specifically through single-class binary labeling.

3.1.3 Balancing the level of difficulty to elicit motivation and improve label quality: We observed that participants became demotivated when the difficulty level was either overly simple or excessively challenging. To achieve a moderate challenge, we carefully adjusted the difficulty level by tuning the parameters of the bot's algorithm that we employed through a trained DCNN model on FER+ [4]. We determined specific parameters for the bot's algorithm based on participant evaluations from the design probes. Additionally, the number of game rounds and time restrictions were also decided in the design probes, all of which are depicted in Figure 2.

The difficulty level was also affected by the social pressure from other players, which could induce a biased response toward one side, hindering the labeling and learning of diverse perceptions. This issue could be resolved by providing equal opportunities to all players. To ensure equal opportunities, we employ two game elements in our final design, which are turns and anonymity. During gameplay, we designed all players to be anonymized without any displayed rankings or badges, and they take turns equally in an order determined randomly at the beginning of each game round.

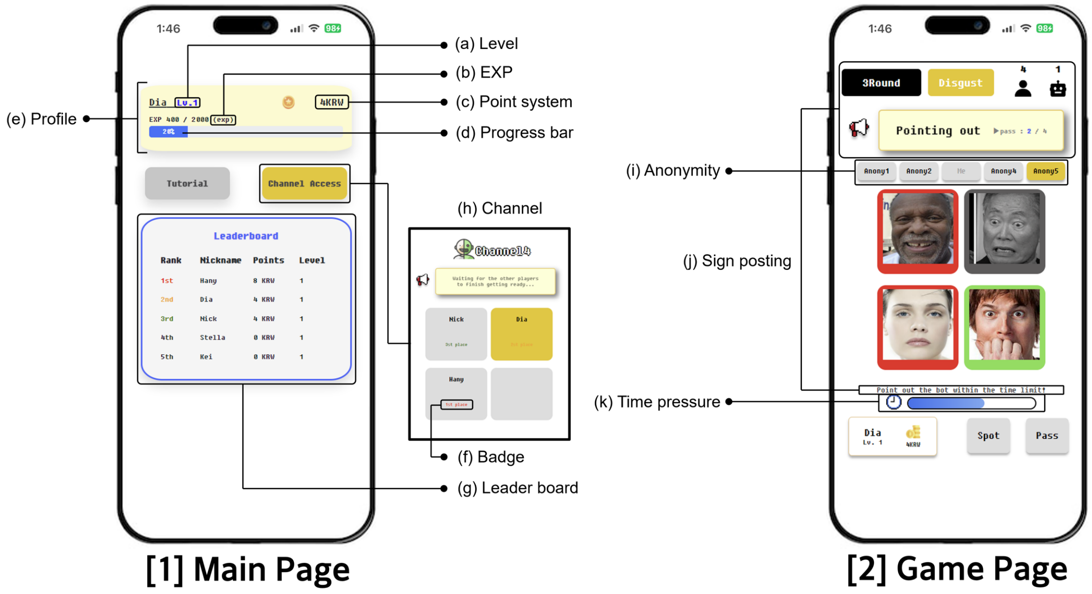
3.1.4 Attractive visual design: Beyond our initial expectations, participants provided substantial feedback on the look and feel of the UI design. We found that an aesthetically pleasing UI design, including a clear layout and consistent style and color themes, is one of the key design elements for successful gamification. Therefore, in the final design, as shown in Figure 2(d), we used a color chart to ensure consistency in the overall theme and included rich visual elements associated with the game's storyline.

4 FIND THE BOT!

Below, we describe our overall game design including gamification strategies and the gameplay that aim to seamlessly incorporate our design goals. To illustrate how our game design supports effective training and reliable label collection while players enjoy the game, we walk through a scenario following Dia, who is playing the game in her free time. We then give a brief overview of the app's implementation.

4.1 Gamification Strategies

To effectively engage and motivate people within an interface, we designed our game using a combination of game elements based on existing gamification strategies [17, 19, 33, 44, 70] and our findings from the design probing. The game elements used in Find the Bot! were designed considering two different levels: an abstract level connected to motivational strategies (first column in Table 8), and a more concrete level that facilitates the implementation of these strategies (second column in Table 8). We came up with 10 motivational strategies, each implemented through 14 concrete level game elements that are suitable for our game. These are summarized in Table 8 in the Appendix A.



Leaderboard

Rank	Nickname	Points	Level
1st	Hany	8 KRW	1
2nd	Dia	4 KRW	1
3rd	Nick	4 KRW	1
4th	Stella	0 KRW	1
5th	Kei	0 KRW	1

Waiting for the other players to finish getting ready...

Point out the bot within the time limit!

Figure 3: In both [1] the Main Page View and [2] the Game Page View, game elements fulfill specific roles. On the main page, players can track their own status and progress in the game, aided by a variety of motivating game elements listed from (a) to (h). On the game page, an array of components assists players in following the game flow and becoming fully immersed in the gameplay. Detailed descriptions of these elements are provided in Table 8 in the Appendix A.

4.2 Gameplay of Find the Bot!

In the classic Mafia game, each player is secretly assigned a role as either Mafia or Innocent and takes turns trying to guess who the Mafia members are based on each player's responses. If players detect suspicious responses, they can vote to deactivate the players who appear to be Mafia after listening to their 'last comment' — a speech that players can make in their own defense. When the number of Mafia members equals or surpasses the number of Innocents, the Mafia achieves an immediate victory.

Find the Bot! is inspired by this traditional game, but has tailored the gameplay to focus on FER training and labeling. As summarized in Table 2, Find the Bot! involves four human players, who are anonymized in the game, along with an AI 'bot' that has slightly lower FER ability than average people. The game goal is for human players to successfully spot the bot among themselves through real-time interaction and cooperation while trying to avoid raising suspicion of being the bot. The game consists of four rounds, each featuring a randomly chosen emotion keyword from the basic emotions (happiness, sadness, surprise, fear, disgust, anger, contempt, and neutral) [22]. In each round, six game stages unfold: labeling, skimming, pointing out, voting, last defense, and advice. The game ends immediately when the win or lose conditions are met, even if it is in the middle of a game stage. We provide detailed descriptions of each game stage and flow, accompanied by screenshots and a flowchart, in the Appendix B.

Feature	Description
Number of Participants	Five players (four human players and one bot pretending to be a human)
Game Goal	Finding the bot among players without being deactivated
Lose Condition	Deactivation of two players or failure to find the bot until the final fourth round
Win Condition	Successfully find the bot before the end of the game
Number of Game Rounds	Four rounds with randomly set basic conditions
Game Stages	Six stages (labeling-skimming-pointing out-voting-last defense-advice)

Table 2: The overview gameplay of Find the Bot!

4.3 Game Scenario

To better understand how Find the Bot! engages people in both reliable label collection and effective human FER training, we describe the active interactions between Dia and other players in the game.

4.3.1 Set-up. Dia frequently experiences challenges due to her limited social awareness. Recently, she signed up for a web game, "Find the Bot!", following a friend's recommendation, as a means to enhance her ability to recognize facial emotional expressions in others. Feeling bored on the bus, she decides to play Find the Bot! on her smartphone. After logging into the game, she first learns the

	Learner group (N=11)	**Ordinary player group (N=37)**	**Control group (N=11)**
Task	Task A	Task A	Task B
Gender	1 Female, 10 Male	16 Female, 21 Male	5 Female, 6 Male
Age(M, SD)	21.64, 2.35	20.92, 2.71	21.10, 2.97
FER ability (min, max, M, SD)	26, 40, 36, 3.90	43, 63, 51.57, 4.88	31, 40, 37.18, 2.40

Table 3: Participants demographics. FER ability was measured using a pre-survey (combined JACFEE and JACNeuF) on a scale from 0 (low ability) to 64 (high ability).

rules through a tutorial. As the rules are based on the famous Mafia game plot, she quickly understands the gameplay. To access a game channel, Dia is directed to the main page (Figure 3(1)) of the game website. On this page, Dia can view her profile (Figure 3(e)), which includes points, level, and a progress bar, as well as the leaderboard (Figure 3(g)). Dia clicks on 'channel access' (Figure 3(h)) to play with other players. Upon entering channel 1, she encounters a fluent player with the nickname 'Hany', who has a high ranking with a red badge (Figure 3(f)), which motivates her to improve her ranking on the leaderboard after playing the game.

4.3.2 *Playing Find the Bot!*

Now, an emotion keyword for the first round, 'contempt', is presented on the screen (Figure 3(2)). All players, including Dia, are now anonymized. Everyone pays close attention to label images as either 'contempt' or not while glancing at the shrinking timer bar (Figure 3(k)). Dia scans her assigned images, and then, in a state of uncertainty, clicks on the image in the top right corner that seems to represent 'contempt'.

After labeling, all players skim each other's labels. During this time, Dia finds an image in the top left corner from 'anony 1' (Figure 3(i)) that is labeled as 'contempt', even though it doesn't seem to represent that emotion. She makes a mental note to point it out as an error on her turn. After a few turns have passed, Dia takes her turn and immediately points out the image from anony 1, identifying it as a bot's error. However, no players agree with her suspicion, thereby invalidating the vote. Feeling puzzled, Dia closely reexamines the image to understand why others see it as representing contempt. Then, next player takes their turn and points out a label from anony 4 as incorrect. While Dia had agreed with anony 4's labels, after voting, the other players do not agree that the image represents 'contempt'.

Having consistently disagreed with others in her emotional judgments, Dia now decides to pay close attention to how other players categorize faces as showing contempt or not. Suddenly, Dia's screen starts flashing red, accompanied by a notification that she had been spotted as a bot! Feeling flustered, Dia quickly scans through her own labels to identify any that might not be convincing to others, but ultimately fails to find the evidence. Dia is now deactivated due to failing in finding the evidence, and she learns how the majority of people interpret the emotions of her given images. After finishing the game, Dia feels a newfound competitive spirit and aims not to get pointed out as a bot and deactivated early in the next game, as she re-enters the game channel.

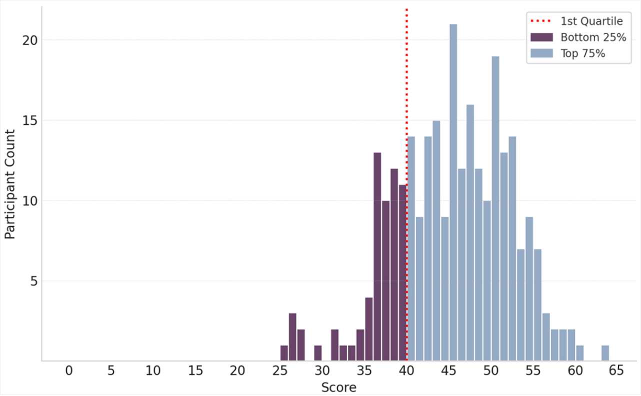


Figure 4: Histogram of FER scores of 275 participants who responded to our online survey in which was aimed to collect ground truth measures for the user study. We used the conventional sign-based scoring. Red dash line indicates the first quartile, which was used as a criteria to group user study participants.

4.4 Implementation

Find the Bot! was implemented with Django web framework, using Python 3.10, HTML5, CSS3, and Javascript. For the back-end, we used MySQL to track user data, including behavior logging and label histories for our user study. Real-time communication and synchronization of game states between users and the server were facilitated using WebSockets and Django Channels. To implement the bot's functionality, we locally trained a state-of-the-art DCNN model based on the VGG13 architecture [57]. This model was trained on the widely used in-the-wild facial expression dataset, FER+ [4], comprising about 35k images categorized into eight emotion classes: neutral, happiness, surprise, sadness, anger, disgust, fear, and contempt. The server was equipped with the trained model to predict emotional labels for images in real-time during the game, effectively serving as the 'bot'. We note that the bot's performance is fixed throughout all gameplay to provide consistent level of difficulty.

5 USER STUDY

To evaluate the feasibility of gamifying and integrating human FER training and FER label collection for machine learning into a single game application, we conducted a user study with Find the

(a) Overview Table of Game Round (in average)

Throughput	
Playtime	390 seconds
Number of collected labels	16
Interaction	
Number of times pointing out others	9.60
Number of times pointed out by others	6.85
Number of Last Mentioning	1.9
Number of Voting	1.07

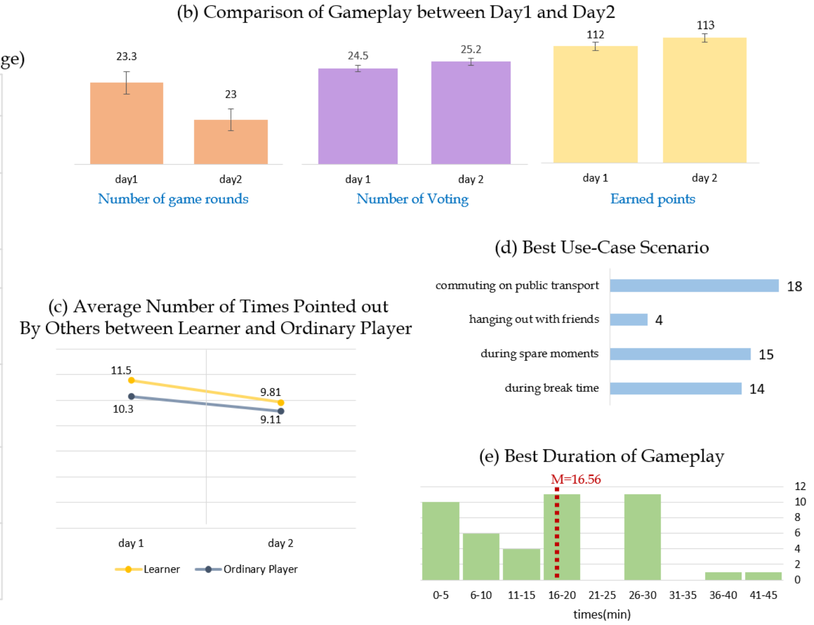


Figure 5: A summary of log data (a-c) and post-survey responses (d-e) from all participants who used Find the Bot! in our user study. Within each game, rich interaction was observed, which is summarized in (a) and (b). As shown in (c), we observed that the number of times being mistaken as a bot decreased in the second day, which implies improved performance in judgment-based FER ability. Participants reported that they would use this game during commuting on public transportation, and the preferred duration of the game as around 16-20 minutes in average.

Bot!. While any in-the-wild dataset could be applied to the game, we used a portion of the AffectNet [58] dataset as spontaneous facial expression images. AffectNet is one of the most widely used large-scale dataset of facial expression images collected in real-world settings, and it includes categories for our target emotions (seven basic emotions [22] and neutral). We randomly selected 28 facial expression images from AffectNet for each emotion category, totaling 224 images used in the study.

We had three main goals for the user study: (1) to assess the quality of overall game design, (2) to assess the effectiveness of judgment-based human FER training, and (3) to assess the quality of collected labels through the game.

5.1 Measures

5.1.1 Assessing the game design: To answer RQ1, we collected and evaluated user experience through the Game Experience Questionnaires (GEQ) [38] and the System Usability Scale (SUS) [9]. We note that we did not employ alternative methods such as the Player Experience Inventory (PXI) [1], as the combination of the GEQ, SUS, and our detailed, customized post-survey would be sufficient to answer our research questions. Moreover, GEQ has multidimensional structure, being widely applicable to various game genres [46]. In addition to these questionnaires, we analyzed participants’ game progress from their log data (summarized in Figure 5) and their responses in a post-survey.

5.1.2 Assessing the effectiveness of FER training: To answer RQ2, we collected ground truth measures from 275 people (aged from 18 to 59) by snowball sampling and online advertising (as shown in Figure 4). The ground truth measures served two purposes: 1) to set criteria to divide our participants into learner and ordinary player groups and 2) to evaluate the judgment-based FER abilities of the learner group before and after the user study. We used Japanese and Caucasian Facial Expressions of Emotion (JACFEE) and Neutral Faces (JACNeuF) [53] as the ground truth measures. These materials, widely used in facial emotion research, consist of eight facial expression images labeled based on AUs. The materials included a total of 64 images (8 for each emotion). We analyzed the distribution of FER scores (M=45.87, SD=6.58) and used the first quartile, which was 40, as a criterion to classify participants into learners who have low FER scores and ordinary players.

5.1.3 Assessing label quality: To answer RQ3, we used Gini coefficent [28] and Fleiss Kappa [26] to measure the level of agreement among all players, which indicates the reliability of the socially agreed-upon emotion labels.

5.2 Participants Recruitment and Study Procedure

We recruited 59 participants (38 male, 21 female; age M=21.08 and SD=2.71) from our university mailing lists and through online advertisements on social media where the recruitment and the experiments were in accordance with our institution’s IRB policies.

Component	Average Score		
	Learners(N=11)	Players(N=37)	All players(N=48)
Positive Affect	2.45(SD=0.33)	2.54(SD=0.39)	2.52(SD=0.36)
Negative Affect	1.59(SD=0.65)	1.44(SD=0.53)	1.47(SD=0.56)
Tension & Annoyance	1.18(SD=0.37)	0.88(SD=0.17)	0.94(SD=0.21)
Competence	2.20(SD=0.48)	2.44(SD=0.24)	2.39(SD=0.28)
Challenge	2.51(SD=0.84)	2.11(SD=0.91)	2.21(SD=0.88)
Flow	2.18(SD=0.83)	2.09(SD=0.71)	2.11(SD=0.73)
Sensory & Imaginative Immersion	2.82(SD=0.32)	2.77(SD=0.28)	2.78(SD=0.26)

Table 4: Component scores of Find the Bot!, as measured by GEQ using a linear scale of 0 to 4. On this scale, 0 represents 'not at all', 1 represents 'slightly', 2 represents 'moderately', 3 represents 'fairly', and 4 represents 'extremely'.

Strategy	Game Element	Response Rate		
		Learners (N=11)	Players (N=37)	All Players (N=48)
Story	Mafia game plot	1.00	0.95	0.96
Challenge	Time pressure	1.00	0.95	0.96
	Difficulty level	0.82	0.92	0.90
Progression	EXP	0.73	0.65	0.67
	Level			
	Progress bar			
Reward	Point system	0.91	0.51	0.60
Competition	Leader board	0.64	0.68	0.67
Status	Badge	0.91	0.68	0.73
Punishment	Deactivation penalty	0.73	0.65	0.67
Reminder	Signposting	1	0.89	0.92
Social pressure	Turn	0.91	0.76	0.79
	Anonymity			
Aesthetic	UI design	0.91	0.97	0.96

Table 5: Evaluation of 14 game elements in Find the Bot! The response rate indicates the proportion of participants (out of N=48) who answered 'yes' to the customized question in the post-survey related to each strategy - game element.

Participants completed a pre-survey to assess their FER abilities score. Based on this score, we classify participants into learners with low FER scores (below 40 scores, total 22 participants) and ordinary player group. Half of the learner group were then assigned to an experiment group (N=11) and the rest half were assigned to a control group (N=11). We detailed participant demographics are summarized in Table 3.

The study was conducted in person and consisted of two different tasks: A) playing Find the Bot! or B) labeling facial expression images. Both the learner and ordinary player groups were assigned to Task A, while the control group was assigned to Task B. We included this control condition, Task B, to serve two purposes: (1) to demonstrate that improved FER scores are result from the use of Find the Bot!, not from getting used to a labeling task or categorical learning and (2) to confirm that the effectiveness of FER training, especially in terms of judgment-based scoring, stems from interacting with a group of people, not from repetitively practicing alone. Task A lasted a maximum of 180 minutes and participants were compensated with 45,000 KRW (approximately 34 USD). Task B lasted a maximum of 90 minutes and participants were compensated with 22,500 KRW (approximately 17 USD). All participants read and signed the informed consent form. Below, we describe the detailed task procedure.

Task A: The study was conducted over two days, with each session lasting 90 minutes. During the first session, participants were provided with an overview of the study (15 mins) and then given direct access to Find the Bot! through their smartphones or personal devices via a URL we shared. Participants were given a tutorial in the Find the Bot! and asked to go through a practice game to familiarize themselves with the interface (15 mins). Then participants were asked to access the game channel and play the game (60 mins). In each channel, a pre-matched team consisting of one learner and three ordinary players entered the game. During this time, we logged all game interactions on the server. During the second session, participants were asked to play the game (60 mins), with a short break of about 5 minutes. After playing the game, participants used an online form to answer a post-survey that included several questions about their experience and measurements of their FER abilities (20 mins).

Task B: The study was conducted for a single day with a maximum duration of 90 minutes depending on the participant's labeling speed. To provide a consistent experience with Task A, we used the same set and amount of facial expression images as were used in the game for Task A. Participants first were explained to an overview of the study (15 mins). Then participants were asked to access an online form and then to label 200 spontaneous facial expression images (45 mins) with eight emotion categories (seven basic emotion and neutral). After the labeling task, participants were asked to complete a post-survey that included only measurements of their FER abilities (10 mins).

6 EVALUATION AND RESULTS

We analyzed the results of our user study to assess (1) the quality of overall game design, (2) Find the Bot!'s effectiveness on judgment-based human FER training, and (3) the quality of collected labels through the game. To ensure a learning effect beyond merely understanding others' thoughts about facial expressions to win the game, we assessed items (2) and (3) through study artifacts derived from both in-game contexts (i.e., labels collected during the game) and contexts beyond the game (i.e., pre- and post-tests with FER measurement).

6.1 Evaluation of Game Design

6.1.1 Statistical results. To understand user experiences of Find the Bot!, we quantitatively analyzed game log data throughout the user study (see Figure 5). In a single game round, the average playtime was 390 seconds and the number of collected labels was 16 in average. Within 390 seconds, a significant number of interactions occurred, stemming from each game stage. We observed consistent patterns of gameplay over the two study days, including the average number of game rounds, the number of times reaching to voting stage (driven by pointing out a controversial label), and the points earned. Meanwhile, we noted a trend where both learner and ordinary player groups were less frequently pointed out by others in the second day. These results suggest a progression in their emotional assumptions.

Based on the feedback from participants, we found the optimal playtime is 16.56 minutes. This is because Find the Bot! induces highly concentrated game interactions, which consequently require substantial mental effort. Thus, participants commented that taking breaks within a relatively short timeframe would be the best use-case scenario.

6.1.2 System usability. To assess the usability of our game interface, we used the SUS. The mean SUS score for Find the Bot! was M=80 (SD=11), on a scale from 0 (worst) to 100 (best). The average SUS scores for the learner group and the ordinary player group were M=77.73 (SD=11.04) and M=81.01 (SD=11.31) respectively, and the independent samples t-test results showed no statistically significant difference between the two groups (p=0.40). According to [9], these scores indicate good and acceptable usability. The results suggest that Find the Bot! is designed to be easily used even though it has various complex game elements.

6.1.3 Game experience. As shown in Table 4, we analyzed GEQ responses, where the independent samples t-test and Cohen's d showed no statistically significant difference between the learner group and the ordinary player group (p>.05, effect size <.2). Participants reported that they enjoyed the game and had an overall positive game experience. In response to a post-survey question "Rate how enjoyable the game was on a 7-point scale", the average score was 4.98 (SD=1.26). The results of GEQ also support this, with the high scores for positive affect (M=2.52, SD=0.36) and low scores for negative affect (M=1.47, SD=0.56) as well as very low scores for tension and annoyance (M=0.94, SD=0.21). The high average score (M=2.78, SD=0.26) for Sensory & Imaginative immersion also indicates that participants were impressed and attracted to Find the Bot!. Competence (M=2.39, SD=0.28) and Challenge (M=2.21, SD=0.88) scores suggest that the difficulty of our game was balanced, allowing participants to feel accomplished and confident during gameplay. The average Flow scores were relatively moderate (M=2.11, SD=0.73), which we supposed that the long user study duration (lasting 60 min with a 5-min break) might influenced.

6.1.4 Game elements evaluation. To further assess the effectiveness of each game element in engaging and motivating users, we asked 14 customized questions in post-survey. For example, regarding the 'Story' strategy, we asked "Did the Mafia game plot of finding the hidden bot among people provide enough enjoyment? (answering 'yes' or 'no')". Based on the results (see Table 5), most of the elements were considered well-designed to motivate users with a response rate of over 70%. Chi-Square test and phi coefficient did not find a significant difference in each element between the learner and ordinary player groups (p>.05, effect size<.2), except for the Point system (p=0.04, effect size=0.29).

Additionally, we found that specific elements — Progression, Competition, and Punishment — were relatively accepted at a moderate level across all players. This may be due to the limited experimental duration (1 hour per day, a total of two days), which was not long enough for some participants to make significant progress. Specifically, the primary response from participants was that the leader board with points and EXP (level) effectively motivated them, while a few participants who had a significant gap in scores compared to those at the leader board felt demotivated. Participants commented that they decided to stop trying to close the ranking gap because the difference was too significant to overcome in just two hours. Similarly, participants who were deactivated too frequently also mentioned that the penalty element lowered their motivation. We expect positive feedback on these elements in natural settings, and for ideal duration per play (around 16 minutes).

6.2 Improvement on Judgment-based FER Scoring

6.2.1 Evaluation of FER ability. Descriptive statistics of participants in the learner group and control group on the FER measurement (JACFEE and JACNeuF) from pre-test to post-test are depicted in Table 6. The difference in pre-survey scores between the two groups (learners or control group) was not statistically significant (p=0.402). We used repeated-measures ANOVA and Cohen's d (as effect size) to examine the significance of pre-to-post test score differences within each group and used Bonferroni corrections to correct p-values for multiple comparisons. To assess the effectiveness of tasks from both AUs and group perception perspectives, we applied two scoring systems: sign-based scoring and judgment-based scoring, respectively. Sign-based scoring is the original scoring system defined by the authors (Matsumoto

	Learner Group (N=11)				Control Group (N=11)			
	pre	post	p	*d**	pre	post	p	*d**
Sign-based scoring	36.00 (3.90)	40.55 (5.23)	0.222	0.866	37.18 (2.40)	42.55 (7.56)	0.057	1.02
Judgment-based scoring	38.55 (3.75)	42.64 (5.92)	0.448	0.780	41.27 (3.44)	44.45 (7.27)	0.289	1.126

Table 6: Descriptive statistics (means and standard deviations) and results of group comparisons (repeated-measures ANOVA) of the FER scores. The results show that Find the Bot! can effectively improve learners' FER abilities in both sign-based and judgment-based scoring compared to the control group. ($^*d$ = cohen's d effect size)

and Ekman [6]), based on standardized facial action units in the FACS system. For judgment-based scoring, we used the results of a pre-survey with 275 participants and adopted the responses that the majority of participants selected as answers for the judgment-based scoring system. While all scores were increased after the gameplay and the labeling task, no statistical significance was observed. We believe that this is because of the small sample size. Therefore, we further analyze the agreement of labels after using Find the Bot! in the next subsection.

	Learner Group (N=11)		Control Group (N=11)	
	pre	post	pre	post
Fleiss' kappa	0.45	0.528	0.464	0.484

Table 7: Inter-rater reliability of pre-test and post-test responses computed using Fleiss' kappa. The learner group who used Find the Bot! showed higher agreement after participating in the gameplay.

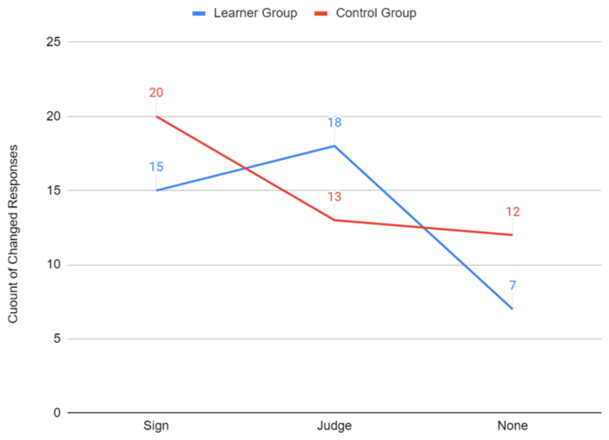


Figure 6: The total number of times that participants changed their post-test judgements on the eight most controversial images. The learner group's responses shifted towards judgment-based answers, while the control group's responses leaned towards sigh-based answers.

6.2.2 Increased Agreement after Using Find the Bot! We analyzed the inter-rater reliability of pre- and post-test responses between the learner and control groups using Fleiss' kappa [26] (see Table 7). The increase in Fleiss' kappa score from the pre- to post-test was larger in the learner group (0.078) compared to the control group (0.02), indicating that Find the Bot! help reach better agreement among responses. We additionally investigated the number of responses that participants altered in their post-test judgments, particularly focusing on controversial facial expression images — 8 out of 64 images (12.5%) that showed different aggregated responses in judgment-based scoring compared to sign-based scoring. As shown in Figure 6, we observed a trend where the learner group's responses shifted towards judgment-based answers, while the control group's responses leaned towards sigh-based answers. The analysis of pre- and post-test responses indicates that the learner group actually learned how the group would perceive facial emotions, extending beyond in-game contexts. Together, these results suggest that Find the Bot! effectively supports judgment-based FER training rather than merely understanding others' thoughts about facial expressions, addressing RQ2.

6.3 Increase on Social Agreement of Collected Labels

In total, we collected 10,193 binary labels (e.g., 'True' for 'happiness' on 'image ID') for 224 spontaneous facial expression images from the AffectNet dataset. To evaluate the quality of labels collected through Find the Bot!, we measured the reliability of labels using the Gini coefficient, a uniformity metric commonly used to evaluate the equality of distributions in economics [28]. Given that we have label distributions generated by multiple users for each image, and each label involves a different number of labelers, the Gini coefficient is considered more suitable for measuring reliability compared to inter-rater reliability measures like Cohen's kappa. We present the result in Figure 7(a).

As shown in Figure 7(a), the solid curve shows the Gini coefficient of label distributions of each image is highly skewed towards 1. More than 90% of images have Gini coefficient > 0.4. This shows that the labels are extremely uneven among eight emotion classes within each facial expression image. These results suggest that Find the Bot! can help annotators produce socially agreed-upon FER labels. Additionally, we measured Gini coefficient for the pre- and post-test results from all players (Figure 7(b)). We observed a shift toward consensus, as evidenced by a more skewed Gini coefficient of responses in the post-test compared to the pre-test. The

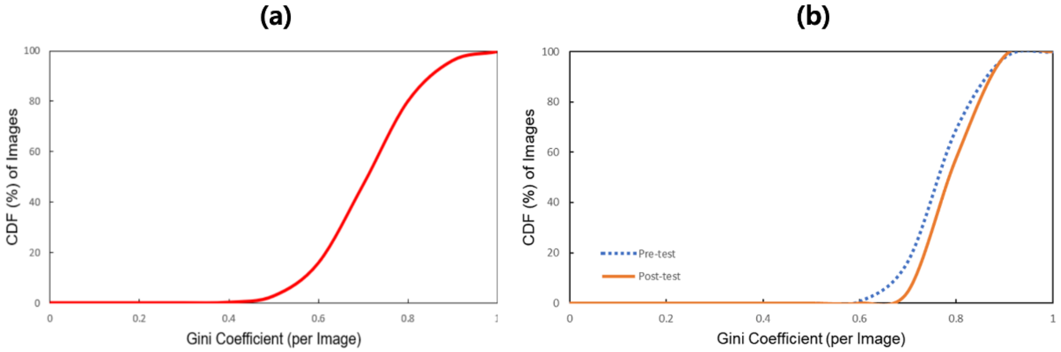


Figure 7: Two graphs depict the cumulative distribution function (CDF) of the Gini coefficient. (a) illustrates the Gini coefficient for collected label distribution throughout the study. (b) displays variations in the Gini coefficient for the distribution of pre-test and post-test results from 48 participants, across 64 facial expression images of our FER measurements (JACFEE and JACNeuF). The blue dashed line represents the distribution of Gini coefficient for the pre-test, and the orange solid line represents that for the post-test. A higher Gini coefficient value indicates more skewed to one emotion label, while a lower value means equal weights for all labels. The significance of the Gini coefficient values are (<0.2 : perfect income equality, 0.2-0.3: relative equality, 0.3-0.4: adequate equality, 0.4-0.5: big income gap, >0.5: severe income gap) [75].

results support that Find the Bot! helps labelers increase the social agreement of the collected labels on spontaneous facial expression images, addressing RQ3.

7 DISCUSSION

We discuss the generalizability, guidelines for game with a purpose, multi-label learning, and possible limitations and future work, reflecting on the lessons learned for this work.

7.1 Generalizability

While we demonstrate the usability of this new gamified interface using an emotional judgment task, we suggest that our findings could be generalized to solve similar but different problems. Specifically, tasks that have the following properties would be especially amenable to our approach:

- The task could be answered in binary (e.g., yes or no) and does not require open-ended responses. Many classification tasks would belong to this class.
- The task can be broken down into the smallest units of work. For example, multi-labeling tasks would not belong to this class because they cannot be reduced to binary labeling of 'true' or 'false'.
- The task is simple enough to allow users to make judgments within a few seconds without a second thought.
- The task embraces subjective responses, but the expected responses should have high agreement rate. For example, movie reviews or satisfaction ratings would not be suitable, because the responses may be dispersed.

We found that many common crowdsourcing problems have these properties, particularly in domains where answers are ambiguous or subjective, such as entity recognition [41], information retrieval [47], or object classification [65]. This is also true for problems in computer-based training of human perception, like judgments of others' thoughts, empathy, personality, and intention from speech or text [7, 81]. We suggest that a range of domains beyond the one explored in this paper may also benefit from our approach.

7.2 Guidelines for Game with a Purpose

In our study, we identified two important considerations in designing game interfaces to successfully engage and motivate participants, especially in controlled environments. We share the insights gained from this process below.

`Benefits from consistent and attractive UI design.` In general, most research-oriented gamified interfaces are implemented using quick-and-dirty methods. Therefore, while attractive UI design is a known element for enriching user experiences, researchers often tend to focus more on other elements or approaches. However, we observed that participants provided positive feedback and expressed satisfaction with Find the Bot!, which includes an appealing UI design as one of our approaches to effective online learning and gamification. In response to the request "Write the best part of this game freely.", majority of participants cited the UI design. P33 and P44 mentioned "The pixel-style graphics, fonts, and animation effects make the game more interesting for me.", and P16 referred "I felt that Find the Bot! could be released as a commercial game, not just a test." Although we did not incur a significant cost (in terms of human effort and financial resources) on UI design, since we only used open-source pixel graphics, icons, and fonts, Find the Bot! was able to elicit engaging game experiences from participants. Moreover, with the advance of generative models, researchers, especially those who are not expert designers, could benefit with minimal burden in their design endeavors.

`Usefulness of utilizing mainstream games.` To develop new and novel gamified interfaces, researchers typically aim to implement games from scratch, with entirely plot and rules. However, contrary to expectations of being clichéd, thirty participants

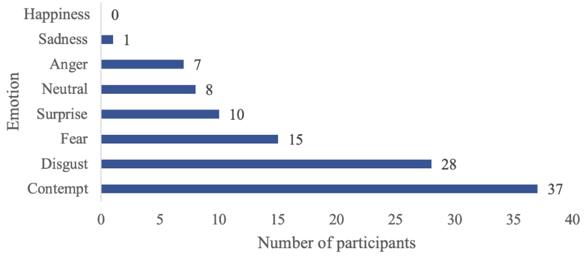


Figure 8: The bar graph displays participants' responses to the question we asked after the user study: "Which emotions did you find difficult to recognize in the facial expression images assigned during gameplay? (Select all that apply)"

(62.5%) cited the Mafia game storyline as one of the most motivating elements. Rather, P18 commented "It's creative to connect facial emotion recognition, a completely unrelated topic, with the plot of the Mafia game." Based on these observations from our study, we noted that using mainstream game plot can benefit the design of gamified interfaces without hindering creativity. Additionally, its familiar rules not only reduce the time participants need to adapt to the system but also make it useful for redesigning to align with specific design goals.

7.3 Multi-Label Learning

In our study, most of the collected labels for each facial expression image were skewed toward a single major emotion, but there were cases where collective annotations were distributed among multiple categories. Detailed examples are provided in Figure 10 in the Appendix C. We found that this was because these images were perceived as conveying compound or ambiguous emotions, or not fitting into any specific basic emotion categories for the participants. Recent previous research has proposed methods for annotating all valid multiple labels in order to robustly train models on datasets with multiple valid interpretations of spontaneous facial expressions captured in real-world settings [4, 49]. Additionally, a line of research has introduced probabilistic approaches for effective multi-label learning by optimally integrating labels and verified the approaches outperform the commonly used majority vote heuristic [49, 84].

Inspired by this literature, it is possible to leverage additional data, such as the results of votes by users in the voting game stage (e.g., three users agreeing with this label or all users disagreeing with this label), as 'implicit labeling'. Integrating this implicit labeling when deciding weight of each category might enhance the quality of labels for spontaneous facial expression images. There also can be an opportunity to more effectively incorporate 'bot' labels (from a DCNN model) in deciding the weight of emotion category by leveraging people's agreement/disagreement with these labels. Future work could verify the feasibility of these approaches.

7.4 Limitations and Future Work

Our work has a couple of limitations. As more complex AI algorithms are being introduced in the context of growing prominence of generative models and large language models (LLMs), simple labeling on static images may not suffice to train these complex models. If wanting to gamify the data collection for these complex tasks, more advanced game design and scenarios may be required.

We only used single static pictures to help improve FER ability of the players, but using a sequence of pictures or even videos may provide better training effectiveness. Future work may use more diverse materials within the game to help arrive at more rich judgment-based agreement among players.

The relatively high complexity of the game may not suit specific groups of people, such as the elderly, even though they are included in non-clinical populations. This restricts the target user group to those who can play digital games and familiar with web environments. To gather more reliable and varied interpretations of spontaneous facial expressions from a wide range of non-clinical populations, future work could focus on tempering the game design without hindering entertainment and engagement.

We did not collect the socioeconomic status (SES) among user study participants, which limits the results to be representative of participants from diverse social backgrounds. Future work may test the proposed game design with more diverse participants to verify whether the findings from this study holds even for the diverse groups.

Participants in the user study use Find the Bot! for a total of 120 minutes over two days. In literature [52], despite variations in training duration across studies (M=6.37 hr, SD=8.34, Min=5 min, Max=35 hr), the authors observed that training duration did not impact training effectiveness. This suggests that training effectiveness is primarily determined by the approach used, particularly a combination of instruction, practice, and feedback. Nonetheless, extending the training duration and then further investigating the training effectiveness of Find the Bot! can be future work.

Some participants noted that game elements related to rewards might need improvement to effectively motivate users. They commented that the current point system can be demotivating in some cases, as it seems unlikely for them to appear on the leaderboard. Thereby, participants suggested an additional element, such as personal rankings, that would allow them to track their relative positions among all players.

As shown in Figure 8, participants reported certain emotions such as contempt, disgust, and fear, particularly challenging to label during the game rounds. It would be beneficial to expose users to these specific emotions more frequently. We primarily focused on exploring the viability of training and data collection using our gamified interface in lab environment, and did not extensively investigate user experience and usage patterns in more diverse situations.

8 CONCLUSION

In this work, we hypothesize that a well-designed game can engage and motivate people — both learners and others — for FER training and labeling using spontaneous facial expression images. In addition, motivated players not only contribute to higher labeling quality but also enhance learning effectiveness (in terms of increasing judgment-based scoring), driven by enjoyment rather than monetary benefits. In our evaluation with 59 participants, we show that using Find the Bot! can increase agreement in judgment-based scoring, which was effective in both training FER for those with low FER scores, and collecting labels with higher social agreement. We identified the effectiveness of specific elements within our game and summarized generalizable insights and recommendations for designing engaging and motivating games with a purpose, which includes using consistent and attractive UI design and utilizing mainstream games. We also suggest generalizable guildelines to be applied to other similar but different tasks.

ACKNOWLEDGMENTS

We thank Juhyung Park for support in development, and Suyeon Shin for design assistance. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (NRF-2022R1C1C1003123), and the DGIST Start-up Fund Program (2021070007).

REFERENCES

[1] Vero Vanden Abeele, Katta Spiel, Lennart Nacke, Daniel Johnson, and Kathrin Gerling. 2020. Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies* 135 (2020), 102370.

[2] C. Ardito, M. De Marsico, R. Lanzilotti, S. Levialdi, T. Roselli, V. Rossano, and M. Tersigni. 2004. Usability of E-Learning Tools. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Gallipoli, Italy) *(AVI '04)*. Association for Computing Machinery, New York, NY, USA, 80–84. https://doi.org/10.1145/989863.989873

[3] Lisa Feldman Barrett. 2017. *How emotions are made: The secret life of the brain.* Pan Macmillan.

[4] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction.* 279–283.

[5] Renae Beaumont and Kate Sofronoff. 2008. A multi-component social skills intervention for children with Asperger syndrome: The Junior Detective Training Program. *Journal of Child Psychology and Psychiatry* 49, 7 (2008), 743–753.

[6] Michael Biehl, David Matsumoto, Paul Ekman, Valerie Hearn, Karl Heider, Tsutomu Kudoh, and Veronica Ton. 1997. Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal behavior* 21 (1997), 3–21.

[7] Danielle Blanch-Hartigan, Susan A Andrzejewski, and Krista M Hill. 2012. The effectiveness of training to improve person perception accuracy: a meta-analysis. *Basic and Applied Social Psychology* 34, 6 (2012), 483–498.

[8] Danielle Bragg, Naomi Caselli, John W. Gallagher, Miriam Goldberg, Courtney J. Oka, and William Thies. 2021. ASL Sea Battle: Gamifying Sign Language Data Collection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 271, 13 pages. https://doi.org/10.1145/3411764.3445416

[9] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996), 189–194.

[10] Andrew R Brown and Bradley D Voltz. 2005. Elements of effective e-learning design. *The International Review of Research in Open and Distributed Learning* 6, 1 (2005).

[11] Hendrik Buimer, Thea Van der Geest, Abdellatif Nemri, Renske Schellens, Richard Van Wezel, and Yan Zhao. 2017. Making Facial Expressions of Emotions Accessible for Visually Impaired Persons. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 331–332. https://doi.org/10.1145/3132525.3134823

[12] K. Byron, S. Terranova, and S. Nowicki. 2007. Nonverbal emotion recognition and salespersons: linking ability to perceived and actual success. *Journal of Applied Social Psychology* 37 (2007), 2600–2619. Issue 11. https://doi.org/10.1111/j.1559-1816.2007.00272.x

[13] John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[14] Ruth C Clark and Richard E Mayer. 2023. *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning.* john Wiley & sons.

[15] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit Players. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.

[16] MM Daniels, E Sarte, and J Dela Cruz. 2019. Students' perception on e-learning: a basis for the development of e-learning framework in higher education institutions. In *IOP Conference Series: Materials Science and Engineering*, Vol. 482. IOP Publishing, 012008.

[17] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From game design elements to gamefulness: defining" gamification". In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments.* 9–15.

[18] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2011. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE international conference on computer vision workshops (ICCV workshops).* IEEE, 2106–2112.

[19] Adriano Lages dos Santos, Mauricio R de A Souza, Eduardo Figueiredo, and Marcella Dayrell. 2018. Game Elements for Learning Programming: A Mapping Study.. In *CSEDU (2).* 89–101.

[20] Andrew Downs and Paul Strand. 2008. Effectiveness of emotion recognition training for young children with developmental delays. *Journal of Early and Intensive Behavior Intervention* 5, 1 (2008), 75.

[21] Roselyne Edwards, Antony Stephen Reid Manstead, and Christopher J Macdonald. 1984. The relationship between children's sociometric status and ability to recognize facial expressions of emotion. *European Journal of Social Psychology* 14, 2 (1984), 235–238.

[22] Paul Ekman. 1973. Cross-cultural studies of facial expression. *Darwin and facial expression: A century of research in review* 169222, 1 (1973).

[23] Paul Ekman. 2003. *Micro expressions training tool.* Emotionsrevealed. com.

[24] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[25] Hillary Anger Elfenbein, Maw Der Foo, Judith White, Hwee Hoon Tan, and Voon Chuan Aik. 2007. Reading your counterpart: The benefit of emotion recognition accuracy for effectiveness in negotiation. *Journal of Nonverbal Behavior* 31 (2007), 205–223.

[26] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[27] Ayub Ghasemian and G Venkatesh Kumar. 2017. Enhancement of emotional empathy through life skills training among adolescents students–a comparative study. *Journal of Psychosocial Research* 12, 1 (2017), 177.

[28] C. Gini. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche. [Fasc. I.].* Tipogr. di P. Cuppini. https://books.google.co.kr/books?id=fqjaBPMxB9kC

[29] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20.* Springer, 117–124.

[30] Judith A Hall, Susan A Andrzejewski, and Jennelle E Yopchick. 2009. Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of nonverbal behavior* 33 (2009), 149–180.

[31] Judith A Hall, Debra L Roter, Danielle C Blanch, and Richard M Frankel. 2009. Nonverbal sensitivity in medical students: Implications for clinical interactions. *Journal of general internal medicine* 24 (2009), 1217–1222.

[32] Judith A Hall, Amy N Ship, Mollie A Ruben, Elizabeth M Curtin, Debra L Roter, Sarah L Clever, C Christopher Smith, and Karen Pounds. 2015. Clinically relevant correlates of accurate perception of patients' thoughts and feelings. *Health communication* 30, 5 (2015), 423–429.

[33] Stuart Hallifax, Audrey Serna, Jean-Charles Marty, Guillaume Lavoué, and Elise Lavoué. 2019. Factors to consider for tailored gamification. In *Proceedings of the annual symposium on computer-human interaction in play.* 559–572.

[34] Madeline B Harms, Alex Martin, and Gregory L Wallace. 2010. Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies. *Neuropsychology review* 20 (2010), 290–322.

[35] Jinni Harrigan, Robert Rosenthal, and Klaus Scherer. 2008. *New handbook of methods in nonverbal behavior research.* Oxford University Press.

[36] Catherine Herba and Mary Phillips. 2004. Annotation: Development of facial expression recognition from childhood to adolescence: Behavioural and neurological perspectives. *Journal of Child Psychology and Psychiatry* 45, 7 (2004), 1185–1198.

[37] Christine Hooker and Sohee Park. 2002. Emotion processing and its relationship to social functioning in schizophrenia patients. *Psychiatry research* 112, 1 (2002), 41–50.

[38] Wijnand A IJsselsteijn, Yvonne AW De Kort, and Karolien Poels. 2013. The game experience questionnaire. (2013).

[39] Carroll Izard, Sarah Fine, David Schultz, Allison Mostow, Brian Ackerman, and Eric Youngstrom. 2001. Emotion knowledge as a predictor of social behavior and academic competence in children at risk. *Psychological science* 12, 1 (2001), 18–23.

[40] Mira Jeong and Byoung Chul Ko. 2018. Driver's facial expression recognition in real-time for safe driving. *Sensors* 18, 12 (2018), 4270.

[41] Sanjay Kairam and Jeffrey Heer. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1637–1648.

[42] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining Human and Machine Intelligence in Large-Scale Crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1* (Valencia, Spain) *(AAMAS '12)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 467–474.

[43] Karl M Kapp. 2012. *The gamification of learning and instruction: game-based methods and strategies for training and education.* John Wiley & Sons.

[44] Karl M Kapp. 2013. *The gamification of learning and instruction fieldbook: Ideas into practice.* John Wiley & Sons.

[45] Kimmy S Kee, Michael F Green, Jim Mintz, and John S Brekke. 2003. Is emotion processing a predictor of functional outcome in schizophrenia? *Schizophrenia bulletin* 29, 3 (2003), 487–497.

[46] Effie L-C Law, Florian Brühlmann, and Elisa D Mekler. 2018. Systematic review and validation of the game experience questionnaire (geq)-implications for citation and reporting practice. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. 257–270.

[47] Matthew Lease and Emine Yilmaz. 2012. Crowdsourcing for information retrieval. In *ACM SIGIR Forum*, Vol. 45. ACM New York, NY, USA, 66–75.

[48] Stephan J Lemmer, Jean Y Song, and Jason J Corso. 2021. Crowdsourcing more effective initializations for single-target trackers through automatic re-querying. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[49] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[50] Christian Aljoscha Lukas, Hugo Trevisi Fuentes, and Matthias Berking. 2019. Smartphone-based emotion recognition skills training for alexithymia-A randomized controlled pilot study. *Internet interventions* 17 (2019), 100250.

[51] Andrej Luneski, Panagiotis D Bamidis, and Madga Hitoglou-Antoniadou. 2008. Affective computing and medical informatics: state of the art in emotion-aware medical applications. *Studies in health technology and informatics* 136 (2008), 517.

[52] Abigail A Marsh and Robert James R Blair. 2008. Deficits in facial affect recognition among antisocial populations: A meta-analysis. *Neuroscience & Biobehavioral Reviews* 32, 3 (2008), 454–465.

[53] David Matsumoto. 1988. Japanese and Caucasian facial expressions of emotion (JACFEE) and neutral faces (JACNeuF). *Intercultual and Emotion Research Laboratory, Department of Psychology* (1988).

[54] Karen McKenzie, Edith Matheson, Kerry McKaskie, Lucie Hamilton, and George C Murray. 2000. Impact of group training on emotion recognition in individuals with a learning disability. *British journal of learning disabilities* 28, 4 (2000), 143–147.

[55] Daniel S Messinger, Leticia Lobo Duvivier, Zachary E Warren, Mohammad Mahoor, Jason Baker, Anne Warlaumont, and Paul Ruvolo. 2015. Affective computing, emotional development, and autism. (2015).

[56] Daniel S Messinger, Leticia Lobo Duvivier, Zachary E Warren, Mohammad Mahoor, Jason Baker, Anne S Warlaumont, and Paul Ruvolo. 2014. Affective computing, emotional development, and autism. (2014).

[57] Microsoft. 2016. FERPlus. https://github.com/Microsoft/FERPlus. GitHub repository.

[58] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.

[59] Ali Mollahosseini, Behzad Hasani, Michelle J Salvador, Hojjat Abdollahi, David Chan, and Mohammad H Mahoor. 2016. Facial expression recognition from world wild web. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 58–65.

[60] Andrew Ng. 2021. MLOps: From model-centric to data-centric AI. *DeepLearning. AI https://www. deeplearning. ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI. pdf* (2021).

[61] Sathiyaprakash Ramdoss, Wendy Machalicek, Mandy Rispoli, Austin Mulloy, Russell Lang, and Mark O'Reilly. 2012. Computer-based interventions to improve social and emotional skills in individuals with autism spectrum disorders: A systematic review. *Developmental neurorehabilitation* 15, 2 (2012), 119–135.

[62] Robert S. Rubin, David C. Munz, and William H. Bommer. 2005. Leading from Within: The Effects of Emotion Recognition and Personality on Transformational Leadership Behavior. *Academy of Management Journal* 48, 5 (2005), 845–858. https://doi.org/10.5465/amj.2005.18803926 arXiv:https://doi.org/10.5465/amj.2005.18803926

[63] Andrew Ryan, Jeffery F Cohn, Simon Lucey, Jason Saragih, Patrick Lucey, Fernando De la Torre, and Adam Rossi. 2009. Automated facial expression recognition system. In *43rd annual 2009 international Carnahan conference on security technology*. IEEE, 172–177.

[64] Katja Schlegel, Ishabel M Vicaria, Derek M Isaacowitz, and Judith A Hall. 2017. Effectiveness of a short audiovisual emotion recognition training program in adults. *Motivation and Emotion* 41 (2017), 646–660.

[65] Viktoriia Sharmanska, Daniel Hernández-Lobato, Jose Miguel Hernandez-Lobato, and Novi Quadrianto. 2016. Ambiguity helps: Classification with disagreements in crowdsourced annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2194–2202.

[66] Henry Silver, Craig Goodman, Gabriela Knoll, and Victoria Isakov. 2004. Brief emotion training improves recognition of facial emotions in chronic schizophrenia. A pilot study. *Psychiatry Research* 128, 2 (2004), 147–154.

[67] Miriam Silver and Peter Oakes. 2001. Evaluation of a new computer intervention to teach people with autism or Asperger syndrome to recognize and predict emotions in others. *Autism* 5, 3 (2001), 299–316.

[68] Peter M Sinclair, Tracey Levett-Jones, Amanda Morris, Ben Carter, Paul N Bennett, and Ashley Kable. 2017. High engagement, high quality: A guiding framework for developing empirically informed asynchronous e-learning programs for health professional educators. *Nursing & Health Sciences* 19, 1 (2017), 126–137.

[69] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*. 254–263.

[70] Raimel Sobrino-Duque, Noelia Martínez-Rojo, Juan Manuel Carrillo-de Gea, Juan José López-Jiménez, Joaquín Nicolás, and José Luis Fernández-Alemán. 2022. Evaluating a gamification proposal for learning usability heuristics: Heureka. *International Journal of Human-Computer Studies* 161 (2022), 102774.

[71] Jean Y Song, Raymond Fok, Alan Lundgard, Fan Yang, Juho Kim, and Walter S Lasecki. 2018. Two tools are better than one: Tool diversity as a means of improving aggregate crowd performance. In *23rd International Conference on Intelligent User Interfaces*. 559–570.

[72] Jean Y Song, Stephan J Lemmer, Michael Xieyang Liu, Shiyan Yan, Juho Kim, Jason J Corso, and Walter S Lasecki. 2019. Popup: reconstructing 3D video using particle filtering to aggregate crowd responses. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 558–569.

[73] Tarik Taleb, Dario Bottazzi, and Nidal Nasser. 2010. A novel middleware solution to improve ubiquitous healthcare systems aided by affective information. *IEEE transactions on information technology in biomedicine* 14, 2 (2010), 335–349.

[74] James W Tanaka, Julie M Wolf, Cheryl Klaiman, Kathleen Koenig, Jeffrey Cockburn, Lauren Herlihy, Carla Brown, Sherin Stahl, Martha D Kaiser, and Robert T Schultz. 2010. Using computerized games to teach face recognition skills to children with autism spectrum disorder: the Let's Face It! program. *Journal of Child Psychology and Psychiatry* 51, 8 (2010), 944–952.

[75] Fei Teng, Jiankun He, Xunzhang Pan, and Chi Zhang. 2011. Metric of carbon equity: carbon Gini index based on historical cumulative emission per capita. *Advances in climate change research* 2, 3 (2011), 134–140.

[76] Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 319–326. https://doi.org/10.1145/985692.985733

[77] Luis Von Ahn, Shiry Ginosar, Mihir Kedia, Ruoran Liu, and Manuel Blum. 2006. Improving accessibility of the web with a computer game. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 79–82.

[78] Luis von Ahn, Mihir Kedia, and Manuel Blum. 2006. Verbosity: A Game for Collecting Common-Sense Facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 75–78. https://doi.org/10.1145/1124772.1124784

[79] Luis Von Ahn, Ruoran Liu, and Manuel Blum. 2006. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 55–64.

[80] Yingqian Wang, Skyler T Hawk, Yulong Tang, Katja Schlegel, and Hong Zou. 2019. Characteristics of emotion recognition ability among primary school children: Relationships with peer status and friendship quality. *Child Indicators Research* 12 (2019), 1369–1388.

[81] Ronald E Warner. 1984. Enhancing teacher affective sensitivity by a videotape program. *The Journal of Educational Research* 77, 6 (1984), 366–368.

[82] Jonathan A Weiss, Kendra Thomson, and Lisa Chan. 2014. A systematic literature review of emotion regulation measurement in individuals with autism spectrum disorder. *Autism Research* 7, 6 (2014), 629–648.

[83] Amy E Wells, Laura M Hunnikin, Daniel P Ash, and Stephanie HM Van Goozen. 2021. Improving emotion recognition is associated with subsequent mental health and well-being in children with severe behavioural problems. *European child &*

adolescent psychiatry 30 (2021), 1769–1777.

[84] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.), Vol. 22. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2009/file/f899139df5e1059396431415e770c6dd-Paper.pdf

[85] Beth T Williams, Kylie M Gray, and Bruce J Tonge. 2012. Teaching emotion recognition skills to young children with autism: a randomised controlled trial of an emotion training programme. *Journal of Child Psychology and Psychiatry* 53, 12 (2012), 1268–1276.

[86] Chih-Hung Wu, Yueh-Min Huang, and Jan-Pan Hwang. 2016. Review of affective computing in education/learning: Trends and challenges. *British Journal of Educational Technology* 47, 6 (2016), 1304–1323.

[87] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2018. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision* 126 (2018), 550–569.

A GAME ELEMENTS OF FIND THE BOT!

Strategy	Game Element	Description
Story	Mafia game plot	Enhancing user adaptability and appeal through the adoption of the popular Mafia game plot.
Challenge	Time pressure	Encouraging users to focus on labeling and reviewing all labels while monitoring a timer, with time limits set for each game stage: 10 seconds, 4 seconds, 10 seconds, 10 seconds, 10 seconds, and 5 seconds (Figure 3(k)).
	Difficulty level	Enabling the achievement of a moderately difficult game goal by adjusting the performance in the algorithm of a bot.
Progression	EXP	Allowing users to track progress by rewarding them with 200 experience points (EXP) after each game, regardless of the win or lose condition (Figure 3(b)).
	Level	Indicating threshold points of experience points - gaining 2000 EXP at each level and automatically leveling up based on their participation (Figure 3(a)).
	Progress bar	Visually indicating the percentage of EXP that has been gained by gradually filling its empty space with solid segments (Figure 3(d)).
Reward	Point system	Rewarding points differentially based on the win/lose conditions and the phase of the game round when the game goal is achieved (Figure 3(c)).
Competition	Leaderboard	A board displaying rankings, nicknames, points, and levels of the top five users, initially sorted by points and then by level comparison. Allowing users to visualize their position compared to other users and motivate them to progress through competition (Figure 3(g)).
Status	Badge	A visible symbol for the top five users, marked with different colors corresponding to their rankings(red, yellow, green, and gray). Boosting user self-efficacy and satisfaction through showing their achievements to others (Figure 3(f)).
Punishment	Deactivation penalty	Watching the remaining gameplay and receiving only half of the points, even when the team wins, as a penalty for being deactivated.
Reminder	Signposting	Guiding actions and ensuring that users enter the game flow through game signposting and brief game captions (Figure 3(j)).
Social pressure	Turn	Sequential participation by taking turns to point out the bot equally.
	Anonymity	Anonymizing all users during gameplay to avoid the influence of specific users (Figure 3(i)).
Aesthetic	UI design	Designing the game interface with a color theme of yellow and blue, featuring pixel-art graphics, fonts, meaningful icons, and animations linked to game events for visual appeal.

Table 8: Description of game elements in Find the Bot! The 14 game elements are grouped based on an overarching motivational strategy, and their equivalents in previous studies [17, 19, 44, 70].

B DETAILED GAMEPLAY WITH SIX STAGES

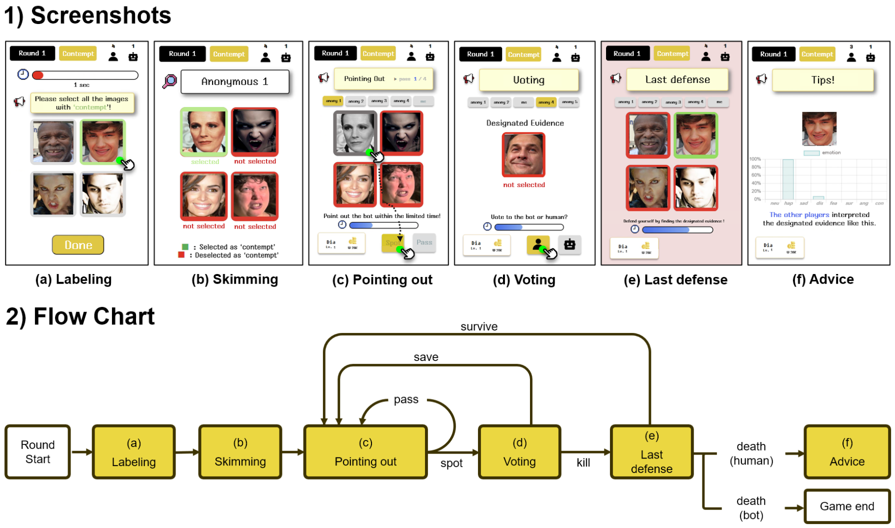


Figure 9: The screenshots and flowchart depict the six game stages in Find the Bot! (1) The screenshots illustrate Dia's game flow, as described in Section 4.2. (2) Throughout the game flow, players progress through various stages, which are determined by evolving game interactions. At the start of the game, players enter (a) the labeling stage. After labeling, (b) players skim through all the labels from each other, before (c) pointing out those suspected of being the bot. Once a player is spotted, the game stage moves on to voting. If the majority agrees with this suspicion, the targeted player is required to present (e) their defense. If the player successfully defends themselves, the game stage reverts to taking turns to point out. Otherwise, the player is deactivated and (f) receives advice. The game continues with players taking turns in this manner until all players pass their turn.

At the start of each round, a randomly set emotion keyword is shown to all players. In the labeling stage (Figure 9(a)), each player, both humans and a bot, is assigned four different facial expression images. They have to label whether each image correctly matches the emotion keyword given at the beginning of the round within ten seconds. For example, if the emotion keyword is 'happiness', players click on all happy-looking images, and those clicked images are labeled as 'True', while others are labeled as 'False'.

After the labeling stage, players move on to the skimming stage (Figure 9(b)), where they can quickly review the generated image-label sets of all players (including the bot) that flick automatically for three seconds per set, totaling twenty image-label pairs (four pairs per player).

In the pointing out stage (Figure 9(c)), every player, except the bot, takes turns to point out a player they believe might be the bot due to suspicious labeling. The turns continue until every player clicks the 'pass' button rather than spotting, indicating that there are no more questionable labels. The player taking his turn has ten seconds to thoroughly inspect labels and spot the bot, while the other players can browse labels in advance to prepare for their turn. For example, if a player sees an image that looks happy but is labeled as 'False', they can click on the labeled image and submit it as 'evidence' of being a bot by clicking the 'spot' button.

Subsequently, all players participate in the voting stage (Figure 9(d)) to conclude whether the pointed player is really a bot or not by evaluating the label's plausibility. If the majority disagrees with the suspicion that the pointed player is a bot, the next player takes his turn and the pointing out stage is repeated. Otherwise, the pointed player enters the last defense stage.

During the last defense stage (Figure 9(e)), the pointed player has five seconds to identify one of the labeled image that may appear unconvincing to the other players, while the others wait for his final defense. If the pointed player successfully identifies the 'evidence', they can survive and the next player takes a turn to point out again. However, if the player fails to find the evidence, they will be deactivated from the game. After a player is deactivated, other players can immediately discover whether it was a player or a bot. The deactivated player then has to watch the rest of the gameplay as a penalty until the game ends.

To help understand the controversial labeled image, especially for the deactivated player, a advice stage (Figure 9(f)) is automatically shown as soon as the player is deactivated. During this stage, all players can view a graph showing the distribution of accumulated emotion labels for the pointed image (the evidence), which has been compiled from all users of Find the Bot! throughout the entire game.

Within the four game rounds, if the human players successfully find the bot by observing each other's labels and silently exchanging votes and feedback, they win. However, if the human players fail to find the bot by the last fourth round, or if two human players are deactivated during the game, they lose. When human players win, they earn both EXP (experience points) and game points, the amount of which depends on how quickly they find a bot. Additionally, players who are deactivated during the game receive half of the game points as a penalty. When human players lose, they receive only EXP but no game points.

C EXCEPTIONAL CASES IN THE DISTRIBUTION OF LABELS IN THE STUDY

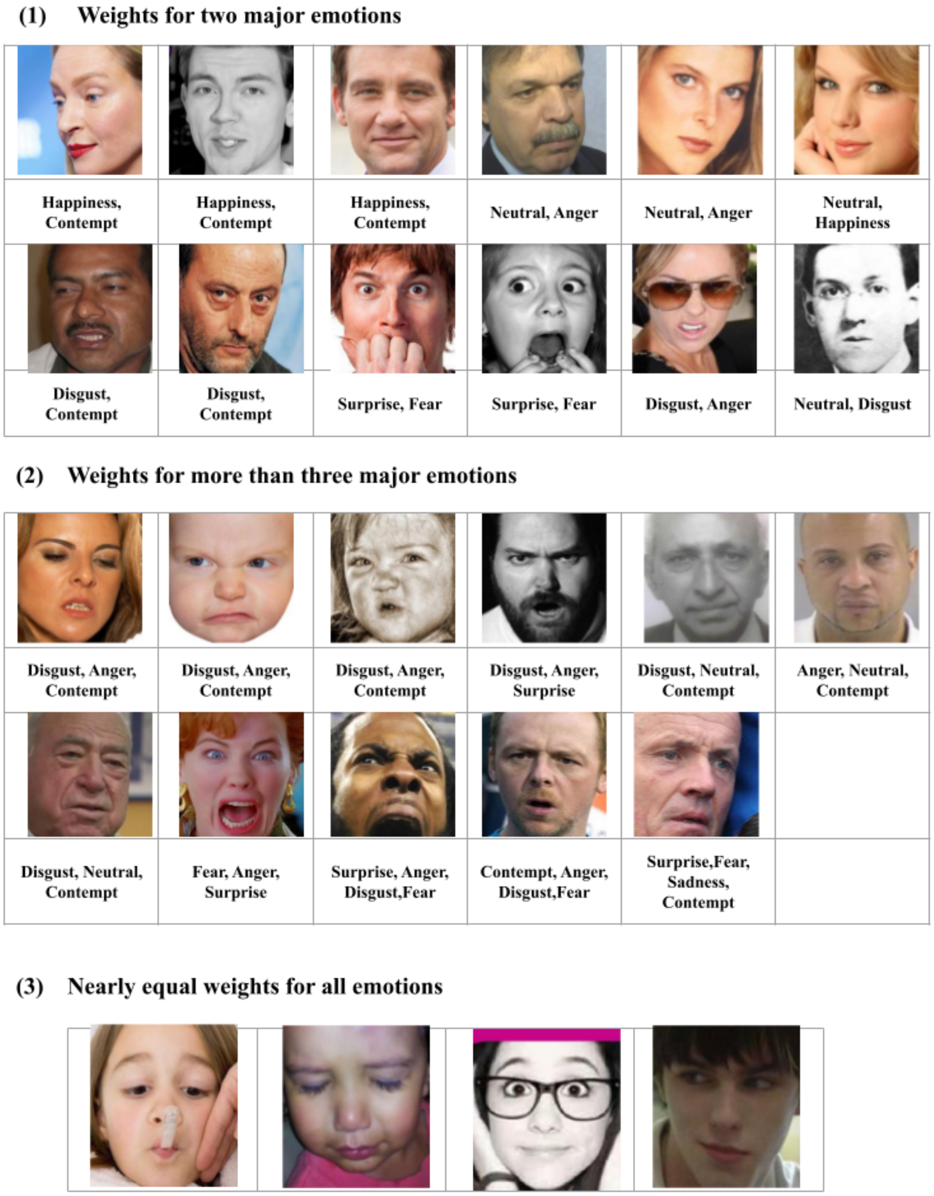


Figure 10: Three exceptional cases in the distribution of labels collected during the study, not skewed toward a single major emotion, are as follows: (1) significant weights assigned to two major emotions, (2) significant weights assigned to more than three major emotions, and (3) nearly equal weights assigned to all emotions.