electronics

MDPI

Article

Intuitive Recognition of a Virtual Agent's Learning State Through Facial Expressions in VR

Wonhyong Lee [1,*] and Dong Hwan Jin [2]

[1] School of Computer Science and Electrical Engineering, Handong Global University, Pohang 37554, Republic of Korea
[2] MyMeta, Seoul 08790, Republic of Korea; daniel@e-mymeta.com
* Correspondence: whlee@handong.edu; Tel.: +82-54-260-1388

Abstract

As artificial intelligence agents become integral to immersive virtual reality environments, their inherent opacity presents a significant challenge to transparent human–agent communication. This study aims to determine if a virtual agent can effectively communicate its learning state to a user through facial expressions, and to empirically validate a set of designed expressions for this purpose. We designed three animated facial expression sequences for a stylized three-dimensional avatar, each corresponding to a distinct learning outcome: clear success (Case A), mixed performance (Case B), and moderate success (Case C). An initial online survey ($n = 93$) first confirmed the general interpretability of these expressions, followed by a main experiment in virtual reality ($n = 30$), where participants identified the agent's state based solely on these visual cues. The results strongly supported our primary hypothesis (H1), with participants achieving a high overall recognition accuracy of approximately 91%. While user background factors did not yield statistically significant differences, observable trends suggest they may be worthy of future investigation. These findings demonstrate that designed facial expressions serve as an effective and intuitive channel for real-time, affective explainable artificial intelligence (affective XAI), contributing a practical, human-centric method for enhancing agent transparency in collaborative virtual environments.

Keywords: virtual reality; facial-expression-based transparency; explainable AI (XAI); human–AI interaction

check for updates

Academic Editors: Deepti Aggarwal and Sophie Mckenzie

Received: 15 May 2025
Revised: 21 June 2025
Accepted: 27 June 2025
Published: 30 June 2025

Citation: Lee, W.; Jin, D.H. Intuitive Recognition of a Virtual Agent's Learning State Through Facial Expressions in VR. *Electronics* **2025**, *14*, 2666. https://doi.org/10.3390/electronics14132666

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, virtual agents and digital systems designed for human collaboration have advanced rapidly, with many major manufacturers developing them [1]. As collaboration between humans and these systems becomes more common, there is increasing interest in equipping them with social communication capabilities, including providing human-like faces to virtual agents for richer interaction [2,3]. In the foreseeable future, people will frequently interact with these types of collaborative and interactive AI agents in immersive environments, such as virtual reality (VR) or augmented reality (AR), for education, training, and services [4].

For these interactions to be effective, users need to understand what the agent is thinking or learning in real time. In other words, the agent's internal state and intentions should be transparent to the user, as this can improve the user's trust and engagement [5,6]. In the field of artificial intelligence, this motivation aligns with the objectives of explainable

AI (XAI), namely making an AI's decision process understandable to humans [7]. Modern deep learning systems often operate as "black boxes" whose internal workings are opaque, leading to a lack of interpretability. This opacity is a well-known problem: users may not trust or effectively cooperate with it if they cannot interpret its behavior [8]. The strategic importance of this field is growing, with recent analyses highlighting the need for transparent and trustworthy AI systems to drive user adoption in 2025 and beyond [9,10]. While existing research has explored various XAI techniques [8] and the importance of agent expressiveness in human–computer interaction [11], the specific question of whether users can intuitively identify an agent's internal learning state—a complex cognitive process distinct from basic emotions—using only facial cues in virtual reality has not been systematically investigated. Therefore, the contribution of this work lies not in the avatar technology itself, but in its novel application for transparent communication and the empirical validation of its effectiveness for this specific purpose.

One intuitive approach to increase the transparency in VR is to leverage the social and emotional cues that humans naturally understand. Research in human–AI interaction suggests that endowing AI agents with socio-emotional attributes can foster trust, empathy, and better collaboration [12]. In particular, giving a virtual agent the ability to express emotions or internal states through facial expressions or other nonverbal cues may help human users read the agent's mind in a natural way [11,13]. Emotions are a universal language: psychological studies have shown that humans instinctively respond to facial expressions, and can even attribute mental states to inanimate agents that exhibit facial cues [14]. Facial expressions are one of the most effective nonverbal communication channels for conveying affect and intention. Paul Ekman's seminal work identified six basic emotions (happiness, sadness, fear, anger, surprise, and disgust) with associated facial expressions that are recognized across cultures [15]. These basic expressions, and combinations can be used to communicate a wide range of feelings and states. Importantly, recent studies in educational technology and human–computer interaction have demonstrated that a virtual agent's emotional expressiveness can influence users' perceptions, motivation, and even learning outcomes. For example, the presence of appropriate facial expressions in a pedagogical agent can positively affect learners' emotional state and engagement [16]. Conversely, an agent with mismatched or fake-looking expressions can induce negative reactions or reduce credibility. This view highlights the value of emotional transparency and appropriate expressiveness in virtual agents.

Building on these insights, it is posited that a virtual agent can provide real-time, transparent feedback about its own learning process by displaying intuitive facial expressions. In other words, as the virtual agent learns (for instance, as an AI algorithm trains or a virtual student avatar attempts a task), it can outwardly show signs of its internal cognitive state of learning–a more nuanced process than expressing basic emotions–such as confidence, confusion, or progress, through a human-like face. If users can correctly interpret these facial cues, they gain an immediate understanding of the agent's status without needing technical metrics or explanations. This approach could make interactions with AI in immersive environments more natural and foster a sense of empathy: the user can feel when the agent is struggling or succeeding, much as a human teacher gauges a student's understanding from their expressions. Such emotional feedback could be especially useful in educational or collaborative scenarios, allowing a human to intervene or adjust their assistance based on the agent's apparent state. It effectively embeds a form of explainability into the agent's behavior, by leveraging affective signals.

The primary contribution of this paper is the empirical investigation into the intuitive recognition of a virtual agent's learning state via its facial expressions within a virtual reality setting. The key question is "Can the user correctly discern how well the agent is

learning by watching the its face?" To explore this, a 3D virtual agent (facial avatar) is designed with a stylized human face that can display various facial expressions. A set of facial expressions is identified to represent different possible learning states of the agent (for example, "I'm confidently understanding," "I'm confused," or "I'm struggling but trying"). These expressions were informed by Ekman's basic emotions and by prior studies on student facial expressions during learning (e.g., a nod and smile might imply the agent has grasped the material, whereas a furrowed brow or yawn might imply difficulty or boredom) [17]. Then expression sequences are created to simulate the agent's facial progression over the course of a learning session under different conditions. In a user study, participants wearing a VR headset observed these facial-expression sequences and attempted to infer the agent's learning state. The following primary hypothesis is tested.

Hypothesis 1 ((H1) Main Hypothesis). *Participants will be able to intuitively and accurately recognize the virtual agent's learning state from its facial expressions.*

In addition, we formulated three exploratory sub-hypotheses to examine the potential influence of user background factors: potential differences based on gender (H2), prior virtual reality experience (H3), and academic major (H4). Our results provide strong empirical evidence for our main hypothesis, supporting the concept of facial expressions as a means of real-time, affective explainable AI (affective XAI). Our results provide novel empirical evidence that this approach is viable, offering a new channel for affective XAI that relies on innate human social perception rather than technical readouts. The novelty lies not in the avatar technology itself, but in its specific application to convey cognitive states and the validation of its effectiveness in enhancing human-centric transparency. Our findings show a high overall accuracy of approximately 91% in recognizing the agent's learning state, supporting the effectiveness of this approach for affective XAI in virtual reality.

The remainder of this paper is organized as follows. Section 2 reviews the background and preliminary work that informs our avatar's expression design purpose, including the role of facial expressions in communication and learning. Section 3 describes the agent's facial expression design and a preliminary validation of their intended meanings. Section 4 details the implementation of the avatar in VR and VR experiment settings, and Section 5 presents the experimental results. In Section 6, the findings, implications for explainable AI (virtual agent) and user trust, and limitations of the study are discussed. Finally, Section 7 concludes the paper.

2. Related Work

In this section, we review the foundational prior work that underpins our study. We first discuss the role of facial expressions as a powerful medium for nonverbal communication, then examine how they have been specifically considered in learning contexts, informing our design choices for the virtual agent.

2.1. Facial Expression as a Communication Medium

Facial expressions have long been studied as a fundamental mode of nonverbal communication. Research in psychology and computer vision has established that certain facial patterns reliably convey emotional states and can be recognized by humans universally. Paul Ekman's pioneering work anatomically analyzed facial muscles and showed that people instinctively interpret combinations of facial muscle movements as specific emotions [18]. Ekman identified six basic emotions—surprise, joy, sadness, fear, anger, and disgust—which correspond to prototypical facial expressions that are culturally invariant [19]. These findings imply that even without words, a face can signal how a person

(or by extension, an agent) feels and whether they are, for instance, confident, confused, or interested. Later work has expanded on how subtle changes in expression (raising an eyebrow, a slight smile, etc.) can convey nuanced cues about cognitive and affective states [20,21]. In human–computer interaction, embodied agents with human-like faces leverage this channel to make interactions more natural. Users tend to attribute emotions and intentions to facial expressions of virtual characters, a phenomenon rooted in our innate tendency adaptation to respond to faces [14]. Because the human face is such a salient and attention-catching stimulus, a virtual agent's face offers a powerful medium to communicate its internal states in a way users intuitively understand.

2.2. Facial Expressions in Learning Contexts

When humans learn or solve problems, they often exhibit telltale facial expressions—a furrowed brow might indicate concentration or confusion, a smile might indicate satisfaction, etc. Prior studies have attempted to interpret learners' facial expressions to gauge their level of understanding or confusion. For example, monitoring students' facial expressions in a virtual learning environment could effectively detect their comprehension levels, as certain positive expressions correlated with understanding while negative or neutral expressions signaled lack of understanding [17]. Another study similarly reported that a student appearing happy or nodding is likely following along well, whereas a student showing fatigue or puzzlement likely is not [22]. These works suggest that facial cues can be mapped to learning states (e.g., "engaged vs. disengaged" or "comprehending vs. confused").

In this paper's approach, essentially this idea is inverted: rather than reading a human learner's face, we control a virtual learner's face to display expressions that correspond to its learning progress. We drew upon the literature above to select a palette of expressions for our virtual agent that would intuitively represent certain learning conditions. Table 1 summarizes the categories of learner emotions we considered (adapted from [23]). It is worth noting that these categories could be further grounded in established theories of academic emotion, such as Pekrun's control-value theory [24], which offers a structured taxonomy of emotions directly related to learning activities and outcomes. We distinguished, for instance, a neutral attentive state, a state of high interest/curiosity, one of enjoyment, one of decreasing interest or boredom, and varying levels of concentration (including signs of confusion or deep focus). By defining these distinct expression categories, we aimed to capture the range of feelings a learning agent might have during training—from confident and engaged to frustrated or disengaged.

Table 1. Example learner emotion categories and their indicative facial expressions (based on prior study [23] and other sources)

Perceived Emotion	Description
Away/Idle	Not present or not paying attention (e.g., blank or absence of expression for a period)
Normal	Neutral, baseline state (no particular emotion)
Interested	Showing interest (e.g., eyes focused, slight eagerness)
Enjoyment	Amused or happy (smiling, positive expression)
Decreasing Interest	Losing interest (e.g., mild frown, gaze wandering
Concentrating (neutral)	Concentration without emotion (focused look)
Concentrating (serious)	Intense focus with seriousness (furrowed brow)
Concentrating (immersed)	Deep absorption with some positive reaction (e.g., determined look, slight nodding)

Using these as guidelines, we designed the virtual agent's facial expressions for three prototypical learning scenarios. We consider a scenario where the agent is training on a task (for example, learning to solve problems using a deep learning model) and its performance can be characterized by two aspects: how well it has memorized the knowledge (training phase) and how well it can apply or generalize that knowledge (test phase). We defined: Case A—the agent has learned well in both memorization and application (an overall successful learning state); Case B—the agent learned well in memorization but its application performance is not enough (such as overfitting); and Case C—the agent's learning is mediocre in both respects (partial understanding). These cases reflect different outcomes of a learning process and likely engender different emotions in a learner. For instance, in Case A the agent might feel confident or happy about its success; in Case B it might feel mixed emotions (satisfied with memorization success but perhaps uncertain or confused due to weaker application); and in Case C the agent might feel slightly disappointed at its moderate performance. We mapped these cases to specific expressions by explicitly linking established learning and emotion theories to our design choices. The rationale for each case is as follows:

- For Case A (clear success), which reflects high control and high positive value in Pekrun's terms [24] and represents successful application in Bloom's taxonomy [25], we selected expressions of joy and interest. This aligns with Ekman's universal expression for happiness [15] and observations of satisfied learners [23].
- For Case B (mixed performance—good memorization, poor application), which could elicit a blend of satisfaction and confusion, we incorporated a puzzled expression. This reflects a more complex cognitive state where the agent possesses 'knowledge' (as per Bloom's taxonomy) but struggles with its 'application,' potentially leading to feelings of uncertainty or frustration [22].
- For Case C (moderate success), indicating lower achievement and an unsatisfactory outcome, we designed an expression with elements of sadness and disappointment to reflect that the performance did not meet expectations.

2.3. Advances in Facial Expression Analysis and 3D Face Modeling

Our research on designing communicative expressions for a virtual agent is informed by, yet distinct from, recent technological advancements in two related fields: automated facial expression recognition (FER) and three-dimensional (3D) face modeling.

The field of FER has made significant strides, with many studies focusing on the automatic detection and classification of human emotions from images and videos. These approaches often leverage deep learning architectures and big data to analyze spontaneous expressions with high accuracy [26]. Concurrently, sophisticated techniques in 3D computer vision now enable the automatic and robust correspondence of facial surfaces, often using landmark-free methods to create high-fidelity digital models of human faces [27].

It is crucial, however, to distinguish the goals of these fields from our own. These advanced methods are primarily focused on the analysis and interpretation of real human faces. In contrast, the contribution of our work lies not in the analysis of spontaneous expressions, but in the design and empirical validation of deliberately crafted, stylized expressions intended to communicate a specific internal cognitive state—namely, learning progress—for an artificial agent. Therefore, while we acknowledge the importance of high-fidelity modeling and automatic recognition, our approach intentionally utilizes a stylized avatar to isolate the research question: can a specific set of designed expressions serve as an effective and intuitive communication channel in virtual reality? This focus on communicative clarity over anatomical realism allows us to directly

test our hypothesis regarding affective explainable AI in a controlled manner. Indeed, our approach aligns with the state-of-the-art research trajectory, as recent work presented in 2025 continues to explore dynamic emotion generation from prompts and the challenges of recognizing human emotion in complex, real-world interactions with chatbots [28].

3. Avatar Expression Design and Preliminary Validation

After defining the target expressions for the three learning-state cases, we implemented them on our facial virtual agent using the NaturalFront 3D Face Animation Plugin within the Unity3D (Unity 2020.3.21 LTS) engine. The avatar's face is represented by a 3D mesh, a structure composed of interconnected vertices (small triangles). To create expressions, this mesh is deformed in real-time. The deformation process is driven by a set of adjustable control points on the face. When a control point is moved, it pulls the surrounding vertices with it. The influence on these linked vertices is calculated using a bell-shaped weighting function, where vertices closer to the control point are moved more significantly, ensuring a smooth and organic deformation.

To create a systematic and reproducible set of expressions, the locations and control logic for these points were configured to align with the facial action parameters defined by the Affectiva SDK. This provided a framework of 16 controllable parameters, including brow raise, brow furrow, eye closure, smile, lip pucker, and jaw drop, in addition to head orientation. The final emotional expressions for Case A, Case B, and Case C were not single static poses, but rather short animated sequences scripted by setting and interpolating the values (typically from 0 to 100) of these parameters over time.

For example, the happy expression for Case A, reflecting its successful learning outcome, was constructed using high values for the 'Smile' and 'Cheek raise' parameters. The puzzled look for Case B, designed to reflect the cognitive dissonance of its mixed learning outcome, combined a 'Brow furrow' with a slight head tilt. Similarly, the sad expression for Case C, in line with its mediocre performance, was constructed primarily by activating the 'Lip Corner Depressor' and raising the inner brows via the 'Brow Raise' parameter. The transitions between different emotional states within a sequence were smoothed using sigmoid interpolation over 1.5 s to appear natural rather than abrupt. Figure 1, Figure 2, and Figure 3 show example expression sequences for Case A, Case B, and Case C respectively.



Figure 1. An example expression sequence for Case A (the agent mastered both memorization and application). The images are screen captures of the facial avatar, which was designed according to the methodology in Section 2 and rendered in the Unity3D engine to display the target expressions.



Figure 2. An example expression sequence for Case B (the agent did well in memorization but not in application). The images are screen captures of the facial avatar, which was designed according to the methodology in Section 2 and rendered in the Unity3D engine to display the target expressions.



Figure 3. An example expression sequence for Case C (the agent achieved moderate results in both). The images are screen captures of the facial avatar, which was designed according to the methodology in Section 2 and rendered in the Unity3D engine to display the target expressions.

To verify that our designed expressions were being interpreted as intended, we first conducted an online survey using pre-recorded, 2D videos of the avatar's three expression sequences (this survey was distinct from the main experiment conducted in a VR setting). In this survey, 93 respondents of various ages viewed these sequences (presented in random order) and were asked to match each sequence to the learning scenario it represented. Participants for this online survey were recruited through university online forums and social media, and constituted a convenience sample intended to provide an initial check on the clarity of the designed expressions before the main, resource-intensive virtual reality experiment. While demographic details beyond age range were not the focus of this preliminary validation, the sample size of $n = 93$ was deemed adequate for a robust initial assessment of three distinct animated sequences. We informed participants of the meaning of each case beforehand (e.g., "Case A: the agent mastered both memorization and application; Case B: the agent did well in memorization but not in application; Case C: the agent achieved moderate results in both") and told them to assume the agent's facial expressions correctly reflect its internal state. The respondents thus had a clear mapping of scenario to expected agent feeling, and they simply had to match the shown facial sequence to one of the cases.

The validation results from the 93 respondents were very encouraging: the expression sequences for Case A, Case B, and Case C were correctly identified with rates of 90%, 78%, and 87%, respectively. In other words, a large majority matched the avatar's expressions to the correct learning scenario. Case B's sequence had the lowest recognition rate (78%), which we anticipated because it involved a more complex blend of emotions not corresponding to a single basic emotion (the agent in Case B might look slightly confused yet somewhat content). In contrast, Case A (primarily joyful expressions) and Case C (which included an obvious sad expression at the end) were based on more clear-cut basic emotions, yielding higher recognition. These results gave us confidence that the virtual

agent's facial expressions were indeed conveying the desired information about its learning state. Figure 4 summarizes the validation survey outcomes, showing the percentage of respondents who identified the sequence as belonging to Case A, Case B, or Case C scenarios, respectively.

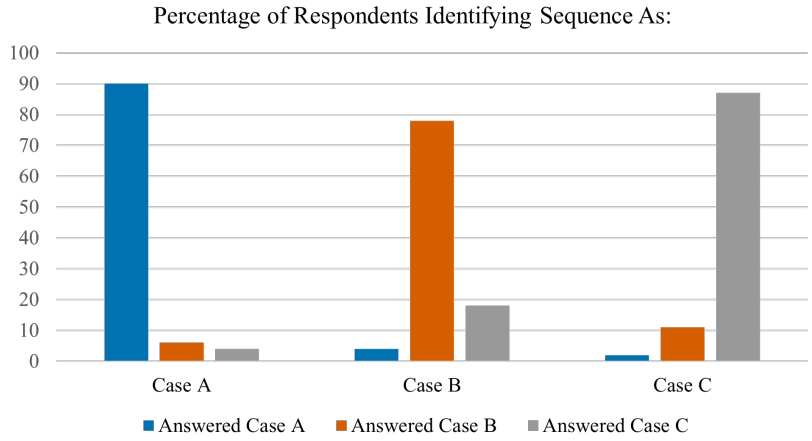


Figure 4. Results of the scenario identification survey ($n = 93$). For each expression sequence presented (e.g., Case A), bars show the percentage of respondents who identified it as belonging to each of the three cases. This chart was created using the graph tool in Microsoft PowerPoint.

4. Experimental Design for Virtual Agent Expression Test in VR

Having validated the expression designs, we proceeded to the main VR experiment to test our hypotheses with users in an immersive scenario.

4.1. Virtual Agent and Apparatus

For the main experiment, an interactive procedure was prepared for participants to observe and interpret the agent's learning-state expressions within a VR environment, using the following setup. We used a head-mounted display (HMD) to present a simple virtual scene containing the 3D agent's face. The agent's face was animated in Unity3D with the preset expression sequences for Case A, Case B, Case C as described in Section 3. Participants could view the agent from a frontal perspective in VR, similar to facing a person. No audio or textual cues were provided—the agent did not speak or display any explicit indicators of its performance, relying solely on facial expressions. The VR scenario was framed as "observe this learning AI avatar and tell us how well it is learning."

4.2. Participants

We recruited 30 participants (university students, 17 male and 13 female, ages 20–26) for the VR experiment. University students were selected for this study as this demographic is relatively more likely to have prior exposure to VR technology, and their academic enrollment allows for a clear distinction between engineering and non-engineering majors for analyzing background influences. To examine these background effects and to investigate potential differences based on gender, we recorded each participant's self-reported VR experience (whether they had used VR before), experience interacting with virtual agents, and academic major. Participants were roughly balanced across these background categories (17 had prior VR experience, 13 did not; 10 were engineering majors, 20 were non-engineering).

4.3. Procedure

To help standardize interpretation and minimize subjective variance, participants were briefed at the start on the meanings of the three learning-state cases (*A*, *B*, *C*) exactly as in

the preliminary survey. We explicitly told them that "the virtual agent's facial expressions will honestly reflect its learning progress." This was to ensure participants approached the task with the expectation that the expressions are informative (no deception). We then explained that the agent would go through a learning session and that they would see a sequence of the agent's facial expressions. Their task was to identify which of the three cases the agent's experience best matched. To avoid order biases, each participant was shown the expression sequences in a random order. That is, the agent would perform three separate learning sessions, and the order of these sessions was randomized per participant. After viewing each sequence, the participant indicated their guess (Case A, Case B or Case C). We included a short break between sequences and ensured that participants understood that each sequence was independent (the agent restarts its learning each time, so there is one of each case per person, in random sequence order). This randomization and independence were important to prevent participants from simply deducing the last case by elimination; each sequence had to be recognized on its own merits.

In our mapping, the agent's memorization ability corresponded to the training accuracy of a deep learning model, and its application ability corresponded to the test accuracy. While participants did not see any numeric accuracies, this conceptual mapping underpinned how we envisioned the agent's internal state: e.g., Case B could be thought of as "high training accuracy (applied knowledge) but medium test accuracy," which might make the agent feel proud of memorizing well but slightly uncertain due to not applying test examples. This internal mapping was not directly disclosed to participants (to avoid confusion with technical terms), but it guided how we animated the subtle cues in expressions. After the VR viewing and responses, we collected basic feedback from participants on whether they found the expressions realistic and any difficulties they faced in interpretation.

5. Experimental Results

5.1. Overall Recognition Accuracy (H1)

Overall, the participants were indeed able to intuitively recognize the virtual agent's learning state from its facial expressions in the VR setting, consistent with our main hypothesis (H1). The aggregate recognition accuracy across all participants and sequences was around 91%. This high success rate corroborates the preliminary online survey findings with an even more controlled immersive test. Figure 5 illustrates the accuracy for each sequence as well as the overall average. The values are also summarized in Table 2. All three learning-state sequences had high recognition significantly above the level of 0.5 ($p < 0.001$), with the sequence representing the mixed-performance case (Case B) again being slightly less consistently identified than the clear-success (Case A) or clear-moderate (Case C) cases. Even for Case B, however, the accuracy was above 90% in the VR study, higher than the 78% in the online survey—perhaps because the immersive 3D avatar gave more depth and dynamic nuance to the expression, aiding interpretation.

Table 2. Mean accuracy ($\hat{p}$), standard error (SE), 95% Wald confidence interval (CI), and binomial test against chance level ($p_0 = 0.5$) for each condition ($n = 30$).

Condition	$\hat{p}$	SE	95% CI	p-Value
Total †	0.91	0.0302	0.85–0.97	<0.001
Case A	0.90	0.0548	0.79–1.00	<0.001
Case B	0.87	0.0614	0.75–0.99	<0.001
Case C	0.97	0.0311	0.91–1.00	<0.001

† Total denotes the mean computed from all 90 trial responses (30 participants × 3 cases).

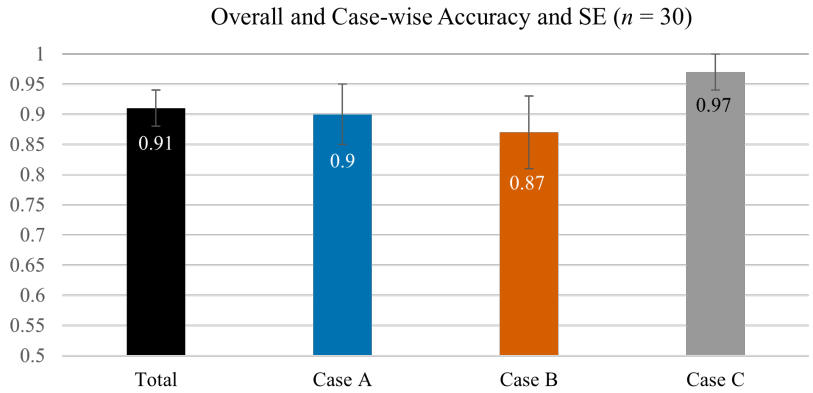


Figure 5. Mean recognition accuracy (proportion correct) for the overall experiment (Total) and for each case ($n = 30$). Error bars represent ± 1 standard error of the mean. This chart was created using the graph tool in Microsoft PowerPoint.

5.2. Subgroup Analyses (H2–H4)

In addition to our main hypothesis, we formulated three exploratory sub-hypotheses to investigate the influence of user background factors:

Hypothesis 2 (H2). *The recognition accuracy of the agent's learning state will differ according to the participant's gender.*

Hypothesis 3 (H3). *The recognition accuracy will differ based on whether the participant has prior virtual reality experience.*

Hypothesis 4 (H4). *The recognition accuracy will differ based on the participant's academic major (engineering vs. non-engineering).*

Because the dependent variable is binary (correct = 1, incorrect = 0), we compared the proportions of correct responses between group pairs using Fisher's exact test (two-tailed), which is well suited for binomial data, especially with small or unbalanced samples [29,30]. For each comparison, we report the exact p-value and the effect size as an odds ratio (OR).

In the following tables, several key metrics are reported to provide a comprehensive analysis. The accuracy, or proportion correct ($\hat{p}$), is the number of correct identifications divided by the total number of trials. The standard error of the proportion (SE), calculated as $\sqrt{\hat{p}(1-\hat{p})/n}$, represents the uncertainty around this sample proportion. The 95% confidence interval (CI), calculated as $\hat{p} \pm 1.96 \times \text{SE}$, provides the range where the true population proportion is expected to lie with 95% confidence. The odds ratio (OR) is a measure of effect size that compares the odds of a correct response in one group to the odds in another. An OR greater than 1 indicates higher odds of success for the first group, while an OR less than 1 suggests higher odds for the second group. The Haldane–Anscombe correction [31] was applied to the cell counts before calculating the OR in cases involving a zero cell.

5.2.1. Gender Effect (H2)

Two-tailed Fisher exact tests compared male ($n_{\text{male}} = 17$) and female ($n_{\text{female}} = 13$) participants for each condition (Table 3, Figure 6). Across all four comparisons, males outperformed females numerically—an effect that was remarkably consistent despite the modest sample size and ceiling-level performance in Case C. Although none of the individual p-values crossed the conventional $\alpha = 0.05$ criterion, the direction of the differences was uniform and the effect sizes were non-trivial: in the pooled analysis (Total) the odds

ratio reached 4.45 with a near-significant trend ($p = 0.073$). This pattern suggests a potential gender-related advantage that the present study was under-powered to confirm statistically. Larger, balanced samples—or a mixed-effects logistic model aggregating data across items and participants—may reveal whether the observed advantage for males reflects a reliable difference in recognising the avatar's facial sequences. At present, however, the evidence remains insufficient to reject the null hypothesis; therefore H2 is provisionally not supported. This implies that while a pattern was observed, we cannot confidently claim a true gender difference based on our data, and both groups demonstrated a high overall ability to interpret the cues.

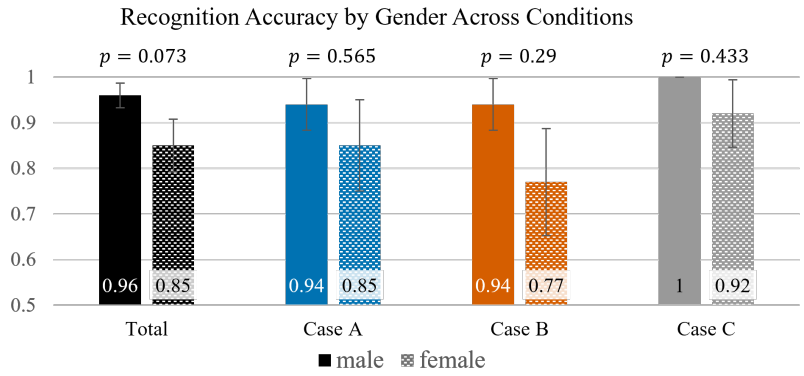


Figure 6. Mean recognition accuracy by gender ($n_{\text{male}} = 17$, $n_{\text{female}} = 13$). Error bars represent ± 1 standard error of the mean, and p-values from Fisher's exact test are displayed above each pair of bars. This chart was created using the graph tool in Microsoft PowerPoint.

Table 3. Accuracy ($\hat{p}$), standard error (SE), 95% Wald confidence interval (CI), odds ratio (OR), and Fisher's exact p-value for the gender comparison in each condition ($n_{\text{male}} = 17$, $n_{\text{female}} = 13$).

Condition	Group	$\hat{p}$	SE	95% CI	OR	p-Value
Total †	Male	0.961	0.027	0.91–1.00	4.45	0.073
	Female	0.846	0.058	0.73–0.96		
Case A	Male	0.941	0.057	0.83–1.00	2.91	0.565
	Female	0.846	0.100	0.65–1.00		
Case B	Male	0.941	0.057	0.83–1.00	4.80	0.290
	Female	0.769	0.117	0.54–1.00		
Case C	Male	1.000	0.000	1.00–1.00	4.20 *	0.433
	Female	0.923	0.074	0.78–1.00		

† Total denotes the weighted mean across the three cases (51 trials for males, 39 for females). * Haldane–Anscombe correction applied where a zero cell occurred [31].

5.2.2. VR-Exp Effect (H3)

Two-tailed Fisher exact tests compared VR-experienced ($n_{\text{exp}} = 17$) and VR-inexperienced ($n_{\text{inexp}} = 13$) participants for each condition (Table 4, Figure 7). In three of the four comparisons (Total, Case A, Case B) the experienced group displayed higher accuracy, and the corresponding effect sizes were substantial: for Case A, where the experienced group made no errors, the Haldane–Anscombe–corrected odds ratio was 11.7 ($p = 0.070$); for the pooled data (Total) the odds ratio reached 4.45 with a near-significant trend ($p = 0.073$). Only Case C deviated from this pattern—here the inexperienced group also achieved perfect accuracy, yielding an odds ratio below 1 (0.41) and a two-sided p-value of 1.00, most likely a ceiling artefact produced by the small number of failures (one in 30 trials).

Although none of the p-values surpassed the conventional $\alpha = 0.05$ criterion, the consistently positive direction of the differences (three of four cases favouring the experienced group) and the non-trivial magnitude of the odds ratios collectively hint at a possible benefit of prior VR exposure. The present study, however, was under-powered to confirm this trend definitively: the groups were unbalanced, several cells contained zero errors, and confidence intervals remained wide. A follow-up with larger, balanced samples—or a mixed-effects logistic model pooling item-level data—may determine whether the observed advantage for VR-experienced users reflects a true performance difference. For now, the evidence is insufficient to reject the null hypothesis, and H3 is therefore provisionally not supported. The practical interpretation is that while prior VR experience appears helpful—as suggested by the strong odds ratio—this study's data does not provide conclusive statistical proof of an advantage.

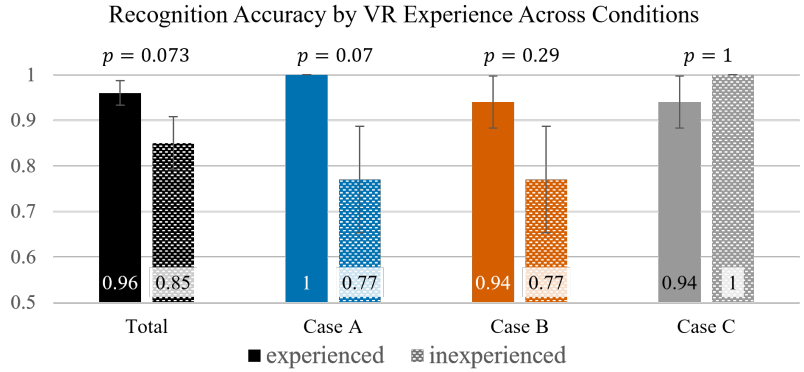


Figure 7. Mean recognition accuracy by virtual reality experience ($n_{\text{experienced}} = 17$, $n_{\text{inexperienced}} = 13$). Error bars represent ± 1 standard error of the mean, and p-values from Fisher's exact test are displayed above each pair of bars. This chart was created using the graph tool in Microsoft PowerPoint.

Table 4. Accuracy ($\hat{p}$), standard error (SE), 95% Wald confidence interval (CI), odds ratio (OR), and Fisher's exact p-value for the VR-experience comparison in each condition ($n_{\text{exp}} = 17$, $n_{\text{inexp}} = 13$).

Condition	Group	$\hat{p}$	SE	95% CI	OR	p-Value
Total †	Experienced	0.961	0.027	0.91–1.00	4.45	0.073
	Inexperienced	0.846	0.58	0.73–0.96		
Case A	Experienced	1.000	0.000	1.00–1.00	11.7 *	0.070
	Inexperienced	0.769	0.117	0.54–1.00		
Case B	Experienced	0.941	0.057	0.83–1.00	4.80	0.290
	Inexperienced	0.769	0.117	0.54–1.00		
Case C	Experienced	0.941	0.057	0.83–1.00	0.41 *	1.000
	Inexperienced	1.000	0.000	1.00–1.00		

† Total denotes the weighted mean across the three cases (51 trials for experienced, 39 for inexperienced). * Haldane–Anscombe correction applied where a zero cell occurred [31].

5.2.3. Major Effect (H4)

Two-tailed Fisher exact tests compared engineering ($n_{\text{eng}} = 10$) and non-engineering ($n_{\text{non}} = 20$) participants for each condition (Table 5, Figure 8). Accuracy was slightly higher for engineers in three of the four comparisons (Total, Case B, Case C), whereas the two groups performed identically in Case A. None of the p-values reached the conventional $\alpha = 0.05$ criterion, but the effect directions were consistently positive and the odds ratios ranged from 1.0 to 1.6. In the pooled analysis the odds ratio was 1.56 and a non-significant

p-value of 0.70. The ceiling-level performance of both groups in Case C (one error in 30 trials) further limited statistical sensitivity.

Although these findings stop short of statistical significance, the uniform pattern in favour of engineering majors hints at a modest advantage that the present study may have been under-powered to detect. A larger, balanced sample—or a mixed-effects logistic model that pools item-level observations—could clarify whether domain expertise contributes meaningfully to recognising the avatar's facial sequences. Given the current data, however, the evidence is insufficient to reject the null hypothesis, and H4 remains provisionally unsupported. The clear interpretation of this null result is that technical background is likely irrelevant for this type of social-cognitive task, a finding that is itself informative.

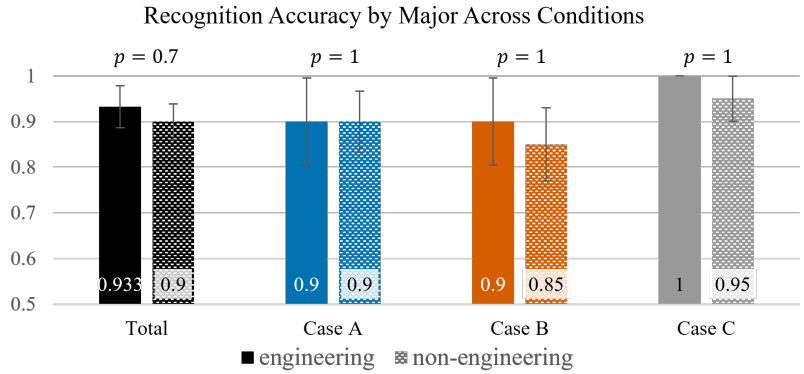


Figure 8. Mean recognition accuracy by academic major ($n_{\text{engineering}} = 10$, $n_{\text{non-engineering}} = 20$). Error bars represent ± 1 standard error of the mean, and p-values from Fisher's exact test are displayed above each pair of bars. This chart was created using the graph tool in Microsoft PowerPoint.

Table 5. Accuracy ($\hat{p}$), standard error (SE), 95% Wald confidence interval (CI), odds ratio (OR), and Fisher's exact p-value for the major comparison ($n_{\text{eng}} = 10$, $n_{\text{non}} = 20$).

Condition	Group	$\hat{p}$	SE	95% CI	OR	p-Value
Total †	Engineering	0.933	0.046	0.84–1.00	1.56	0.70
	Non-eng.	0.900	0.039	0.82–0.98		
Case A	Engineering	0.900	0.095	0.71–1.00	1.00	1.00
	Non-eng.	0.900	0.067	0.77–1.00		
Case B	Engineering	0.900	0.095	0.71–1.00	1.59	1.00
	Non-eng.	0.850	0.080	0.69–1.01		
Case C	Engineering	1.000	0.000	1.00–1.00	1.62 *	1.00
	Non-eng.	0.950	0.049	0.85–1.05		

† Total denotes the weighted mean across the three cases (30 trials for engineering, 60 for non-engineering). * Haldane–Anscombe correction applied where a zero cell occurred [31].

5.2.4. Subgroup Summary and Design Implications

Across all subgroup analyses the ability to read the agent's facial cues remained robustly high (overall accuracy $\approx 91\%$), yet small, non-significant trends did emerge. Males, VR-experienced users, and engineering majors each showed numerically higher accuracies than their comparison groups. None of these trends survived formal significance testing at $\alpha = 0.05$, but the direction and magnitude of the odds ratios (e.g., $\text{OR}_{\text{gender}} = 4.45$, $\text{OR}_{\text{VR}} = 4.45$) hint that certain backgrounds might confer a slight advantage. These observations suggest that future systems could offer adaptive scaffolding: for instance, short familiarisation runs for VR novices or more explicit facial cues for users less accustomed to virtual facial agents.

A recurring pattern was that errors—when they occurred—clustered around the "mixed-performance" sequence (Case B). This sequence blends satisfaction with mild puzzlement and was misclassified most often as the moderate-performance sequence (Case C). Post-experiment interviews reinforced this explanation: several participants remarked that one clip "felt between the other two," and one noted, "When the avatar looked confused, I assumed it did well in one part but not fully." Such feedback suggests that a sharper distinguishing cue—e.g., a fleeting proud smile followed by a puzzled brow—could make Case B more salient without compromising naturalness.

Importantly, no participant regarded the expressions as irrelevant or misleading; most "followed the face to decide." Taken together, the results confirm H1—users can intuitively recognise the virtual agent's learning state—while provisionally leaving H2–H4 (and the exploratory avatar-experience factor) unsupported. The variability observed nevertheless provides actionable insight for tailoring future agents. Figures 6–8 depict the breakdowns for gender, VR experience, and academic major, respectively, and Section 6 discusses how these findings can guide adaptive interface design and cross-cultural calibration.

6. Discussion

Our findings provide a proof-of-concept that a virtual agent's facial expressions can serve as an intuitive, real-time window onto its internal learning state. Participants—without any analytic briefing—correctly inferred how well the agent was learning simply by "reading the face," achieving an overall accuracy of roughly 91%. Figure 5 and Table 2 show that all three learning-state sequences were recognised well above chance ($p < 0.001$). As in the preliminary on-line survey, the mixed-performance sequence (Case B) remained the most challenging, yet even there accuracy exceeded 90 % in VR—substantially higher than the 78% observed on-line—suggesting that the immersive, animated avatar provided additional perceptual nuance that aided interpretation.

6.1. Emotional Transparency as Explainability

Our findings make a specific contribution to the field of explainable AI by providing novel empirical evidence for the viability of affective XAI in virtual reality. The contribution lies not in the avatar technology itself, but in demonstrating that its application can successfully convey a nuanced, internal cognitive state (learning progress), which is distinct from expressing basic, general emotions. The ease with which users decoded the avatar's facial feedback supports our central thesis: an AI agent can explain its status by showing how it feels about its performance. This leverages the user's innate social-cognitive skills and complements existing XAI techniques by offering a continuous, human-readable cue rather than a static technical explanation. Our results thus extend prior claims about socio-emotional transparency to immersive virtual contexts by validating a new communication channel for human–agent interaction.

6.2. Influence of User Factors (H2–H4)

Three exploratory background variables were investigated (Table 3–5). (i) Gender. Males outperformed females numerically in every condition ($OR_{total} = 4.45$), but none of the Fisher p-values reached $\alpha = 0.05$; the strongest trend was $p = 0.073$. A possible reason is that males often adapt more quickly to fully-immersive 3-D environments and outperform females on spatial-orientation tasks in VR [32]. Such spatial ease could free cognitive resources for reading subtle facial cues, which may explain why the biggest male–female gap appeared in the mixed-state sequence (Case B), the most ambiguous of the three. (ii) VR experience. VR-experienced users showed higher accuracy in three of four comparisons, with $OR_{total} = 4.45$ and $p = 0.073$, yet again fell short of significance. Prior

exposure likely reduces sensory and interface load, allowing users to concentrate on the avatar's expression rather than on basic locomotion or headset discomfort. (iii) Academic major. Engineering majors were marginally ahead ($OR_{total} = 1.56$, $p = 0.70$), a direction consistent with evidence that engineering curricula frequently integrate VR tools and thus cultivate interface familiarity [33,34]. However, the very small gap (and $p = 0.70$) suggests that recognising facial emotion is primarily a social–cognitive skill: domain knowledge adds little once the interface itself is mastered.

Across all factors, the directions were consistent but the sample—small and unbalanced, with several zero-error cells—was under-powered to confirm statistical reliability. Practically, these trends imply that minimal acclimation (e.g., a short VR familiarisation or an introductory demonstration) might further flatten the small performance gaps observed.

6.3. Design Lessons from the Mixed-State Sequence

Errors clustered on Case B, whose facial script blended satisfaction with mild puzzlement. This is not surprising: laboratory work on "compound" emotions shows that recognition accuracy drops markedly when two affects are expressed simultaneously, compared with the six basic emotions [35]. Post-session interviews echoed this difficulty—several participants "felt the clip sat between the other two," often mislabelling it as the moderate-performance case. A well-established remedy is to layer additional channels: human observers integrate facial and bodily cues when disambiguating affect [36]; even a quick self-shrug or head-tilt can boost recognition rates. Likewise, giving a robot or avatar a short "inner-speech" utterance (e.g., "Hmm…I'm not sure I got that all correct") has been shown to increase users' understanding and trust, essentially making the agent's thought process explicit. Future iterations could therefore augment Case B with a brief puzzled shrug or a one-line inner-speech cue, or end the sequence with a distinctive relief smile once the agent resolves its uncertainty, to sharpen users' interpretation of this mixed state. This direction is consistent with current trends in multi-modal emotion analysis and empathetic computing, which seek to build more sophisticated and context-aware agents by integrating various communication channels [37].

6.4. Trust and Engagement Potential

Although we did not measure trust directly, participants frequently described the avatar as "alive" or "easy to empathise with." Anecdotally, one user remarked: "When it smiled, I felt proud of it." Such comments echo the "machine-as-teammate" literature, suggesting that facial transparency may foster rapport and sustained engagement.

6.5. Difficulties, Limitations, and Future Work

In conducting this research, we faced several challenges and disappointments that are important to acknowledge. A key design challenge was crafting the "mixed-performance" (Case B) expression. Because it inherently combines conflicting emotional cues (satisfaction and puzzlement), designing a sequence that was both distinct and interpretable was difficult. The lower (though still high) recognition rate for Case B reflects this difficulty. Furthermore, recruitment for in-person virtual reality studies often faces logistical constraints, which in our case led to a relatively small and homogeneous student sample. Achieving broader demographic representation was a practical disappointment that defines a crucial area for improvement.

These challenges inform the study's primary limitations. The most significant constraint, is the small ($n = 30$) and culturally homogeneous student sample. This limits the statistical power of our subgroup analyses (H2–H4) and restricts the generalizability of our findings. Other limitations include the use of a single, stylized avatar, which means our findings may not generalize to other avatar designs; the subjective nature of interpretation,

which can vary by individual and cultural background; and the discretization of learning into only three static states, which is a simplification of real-world dynamic learning.

These limitations, in turn, define a clear path for future research. Future work must aim to replicate and extend our findings with larger, more diverse, and cross-cultural samples. Furthermore, the set of three learning outcomes, while distinct, is limited. Subsequent studies should aim to expand this repertoire to capture a broader spectrum of learner performance and affective states, such as varying degrees of partial understanding, frustration, boredom, or moments of insight. A key avenue for future research is to explore the representation of dynamic learning outcomes that change over time. The current study focused on the recognition of discrete, static learning endpoints. Investigating how an agent's facial expressions can convey an evolving learning trajectory—including transitions between states of understanding, confusion, and insight within a single continuous session—would significantly enhance the ecological validity and utility of this affective XAI approach. Finally, to move beyond subjective user reports, future iterations could integrate objective physiological measures, such as electroencephalography (EEG) or eye-tracking, to provide a valuable complementary channel for understanding the user's cognitive and affective response. A particularly promising avenue for future work is the application of this affective agent framework to industrial-scale technical training. For example, our industrial partner, MyMeta, has been developing a VR-based training suite for semiconductor back-end processing, incorporating emotionally expressive agents into simulations of die bonding and wire bonding procedures. Initial results from this collaboration suggest that learners can interpret facial feedback to assess their procedural understanding, indicating the viability of such systems in specialized engineering education.

Furthermore, we acknowledge that the interpretation of facial expressions inherently involves a degree of subjectivity, influenced by individual differences in perception, cultural background—a factor our homogeneous sample did not allow us to explore—and familiarity with virtual agents. While the high overall accuracy in both our preliminary validation and the main VR study suggests our designed expressions were generally well-understood within our sample, we recognize this as a potential margin of error. Future research should therefore systematically explore these individual and cultural differences to enhance the generalizability of the findings.

7. Conclusions

Our study demonstrates that a virtual agent can make its learning state intuitively transparent by expressing context-appropriate facial emotions in VR. Participants, after only a brief primer, recognised the agent's three learning scenarios with an overall accuracy of 91%, thereby confirming H1. The subgroup analyses (H2–H4) revealed no statistically reliable effects of gender, VR familiarity, or academic major, yet all three factors showed numerically positive—and therefore encouraging—trends. These trends motivate larger, balanced samples and mixed-effects logistic modelling to test whether modest background advantages become significant at scale.

Beyond hypothesis testing, the results endorse facial affect as a lightweight, real-time channel for affective XAI: users could "read" the AI's internal confidence or confusion without needing textual or numeric explanations. Such emotional transparency fits calls in the human–AI interaction literature for socio-emotional cues that foster trust and smooth collaboration.

In sum, giving an VR AI avatar a human-readable "face" offers a practical route to human-centric transparency: users can gauge at a glance whether the system is thriving or struggling and decide when to intervene. As VR and AI grows in ubiquity and complexity, such intuitive feedback may prove vital for keeping humans in the loop and

cultivating empathy-based trust. We regard the present work as a step toward VR and AI systems that not only perform but also communicate their learning journeys in ways people naturally understand.

Author Contributions: Conceptualization, W.L.; methodology, W.L.; investigation, W.L.; writing—original draft preparation, W.L.; supervision, W.L.; data curation, D.H.J.; formal analysis, D.H.J.; visualization, D.H.J.; writing—review and editing, D.H.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by MyMeta and the 2025 Research Year Program of Handong Global University (HGU-2025).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Handong Global University (protocol code 2024-HGUA012 and approved on 27 Jun 2024). for studies involving humans.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: We would like to acknowledge that the 3D virtual facial avatar used in this study was developed with support received during the author Wonhyong Lee's postdoctoral research fellowship in the ART-Med Laboratory, Department of Biomedical Engineering, at The George Washington University, under the guidance of Chung Hyuk Park. We also wish to express our sincere gratitude to Jong Gwun Chong, Jae Hyeok Choi, and Woo Jin Lee for their invaluable assistance with the experimental procedures.

Conflicts of Interest: Author Dong Hwan Jin was employed by MyMeta. The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. Author Wonhyong Lee declares no conflict of interest.

References

1. O'neill, T.; McNeese, N.; Barron, A.; Schelble, B. Human–autonomy teaming: A review and analysis of the empirical literature. *Hum. Factors* **2022**, *64*, 904–938. [CrossRef] [PubMed]
2. Cassell, J. Embodied conversational agents: Representation and intelligence in user interfaces. *AI Mag.* **2001**, *22*, 67.
3. Mayer, R.E. Principles based on social cues in multimedia learning: Personalization, voice, image, and embodiment principles. *Camb. Handb. Multimed. Learn.* **2014**, *16*, 45.
4. Pellas, N.; Mystakidis, S.; Kazanidis, I. Immersive Virtual Reality in K-12 and Higher Education: A systematic review of the last decade scientific literature. *Virtual Real.* **2021**, *25*, 835–861. [CrossRef]
5. Anayat, S.; Kaushik, A. To augment or to automate: Impact of anthropomorphism on users' choice of decision delegation to AI-powered agents. *Behav. Inf. Technol.* **2025**, 1–19. [CrossRef]
6. Daronnat, S.; Azzopardi, L.; Halvey, M.; Dubiel, M. Inferring trust from users' behaviours; agents' predictability positively affects trust, task performance and cognitive load in human-agent real-time collaboration. *Front. Robot. AI* **2021**, *8*, 642201. [CrossRef]
7. Silva, A.; Schrum, M.; Hedlund-Botti, E.; Gopalan, N.; Gombolay, M. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *Int. J. Hum.-Interact.* **2023**, *39*, 1390–1404. [CrossRef]
8. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
9. AlgoAnalytics. The Rise of Explainable AI (XAI): A Critical Trend for 2025 and Beyond. 2025. Available online: https://www.algoanalytics.com/blog/the-rise-of-explainable-ai-xai/ (accessed on 21 June 2025).
10. Riveron. Will 2025 Be the Year of Explainable AI (XAI)? 2024. Available online: https://www.riveron.com/insights/will-2025-be-the-year-of-explainable-ai-xai/ (accessed on 21 June 2025).
11. Pelachaud, C. Modelling multimodal expression of emotion in a virtual agent. *Philos. Trans. R. Soc. B Biol. Sci.* **2009**, *364*, 3539–3548. [CrossRef]
12. Kolomaznik, M.; Petrik, V.; Slama, M.; Jurik, V. The role of socio-emotional attributes in enhancing human-AI collaboration. *Front. Psychol.* **2024**, *15*, 1369957. [CrossRef]

13. Serfaty, G.J.; Barnard, V.O.; Salisbury, J.P. Generative facial expressions and eye gaze behavior from prompts for multi-human-robot interaction. In Proceedings of the Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco, CA, USA, 29 October–1 November 2023; pp. 1–3.
14. Barrett, L.F.; Mesquita, B.; Ochsner, K.N.; Gross, J.J. The experience of emotion. *Annu. Rev. Psychol.* **2007**, *58*, 373–403. [CrossRef] [PubMed]
15. Ekman, P. Facial expression and emotion. *Am. Psychol.* **1993**, *48*, 384. [CrossRef]
16. Wang, Y.; Feng, X.; Guo, J.; Gong, S.; Wu, Y.; Wang, J. Benefits of affective pedagogical agents in multimedia instruction. *Front. Psychol.* **2022**, *12*, 797236. [CrossRef]
17. Sathik, M.; Jonathan, S.G. Effect of facial expressions on student's comprehension recognition in virtual educational environments. *SpringerPlus* **2013**, *2*, 455. [CrossRef] [PubMed]
18. Ekman, P.; Friesen, W.V. Facial action coding system. *Environ. Psychol. Nonverbal Behav.* **1978** . [CrossRef]
19. Ekman, P.; Sorenson, E.R.; Friesen, W.V. Pan-cultural elements in facial displays of emotion. *Science* **1969**, *164*, 86–88. [CrossRef]
20. Keltner, D.; Ekman, P.; Gonzaga, G.; Beer, J. *Facial Expression of Emotion*; Guilford Publications: New York, NY, USA, 2000; pp. 236–248.
21. Ekman, P.; Revealed, E. *Recognizing Faces and Feelings to Improve Communication and Emotional Life*; Emotions revealed; Times Books: New York, NY, USA, 2007.
22. Liu, Y.; Wang, L.; Li, W. Emotion analysis based on facial expression recognition in virtual learning environment. *Int. J. Comput. Commun. Eng.* **2017**, *6*, 49. [CrossRef]
23. Park, J.H.; Jeong, S.M.; Lee, W.B.; Song, K.S. Analyzing facial expression of a learner in e-Learning system. In *Proceedings of the Korea Contents Association Conference*; The Korea Contents Association: Naju, Republic of Korea, 2006; pp. 160–163.
24. Pekrun, R. The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educ. Psychol. Rev.* **2006**, *18*, 315–341. [CrossRef]
25. Bloom, B.S.; Engelhart, M.D.; Furst, E.J.; Hill, W.H.; Krathwohl, D.R. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook 1: Cognitive Domain*; Longman: New York, NY, USA, 1956.
26. Deng, S. Face expression image detection and recognition based on big data technology. *Int. J. Intell. Netw.* **2023**, *4*, 218–223. [CrossRef]
27. Fan, Z.; Hu, X.; Chen, C.; Wang, X.; Peng, S. A landmark-free approach for automatic, dense and robust correspondence of 3D faces. *Pattern Recognit.* **2023**, *133*, 108971. [CrossRef]
28. Kovacevic, N.; Holz, C.; Gross, M.; Wampfler, R. On Multimodal Emotion Recognition for Human-Chatbot Interaction in the Wild. In Proceedings of the 26th International Conference on Multimodal Interaction, San Jose, Costa Rica, 4–8 November 2024; pp. 12–21.
29. Fisher, R.A. Statistical methods for research workers. In *Breakthroughs in Statistics: Methodology and Distribution*; Springer: Berlin/Heidelberg, Germany, 1970; pp. 66–70.
30. Agresti, A.; Kateri, M. Categorical data analysis. In *International Encyclopedia of Statistical Science*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 206–208.
31. Haldane, J. The mean and variance of | chi 2, when used as a test of homogeneity, when expectations are small. *Biometrika* **1940**, *31*, 346–355. [CrossRef]
32. Jacobs, O.L.; Andrinopoulos, K.; Steeves, J.K.; Kingstone, A. Sex differences persist in visuospatial mental rotation under 3D VR conditions. *PLoS ONE* **2024**, *19*, e0314270. [CrossRef]
33. Bano, F.; Alomar, M.A.; Alotaibi, F.M.; Serbaya, S.H.; Rizwan, A.; Hasan, F. Leveraging virtual reality in engineering education to optimize manufacturing sustainability in industry 4.0. *Sustainability* **2024**, *16*, 7927. [CrossRef]
34. Ghazali, A.K.; Ab. Aziz, N.A.; Ab. Aziz, K.; Tse Kian, N. The usage of virtual reality in engineering education. *Cogent Educ.* **2024**, *11*, 2319441. [CrossRef]
35. Jack, R.E.; Garrod, O.G.; Yu, H.; Caldara, R.; Schyns, P.G. Facial expressions of emotion are not culturally universal. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 7241–7244. [CrossRef] [PubMed]
36. Castellano, G.; Kessous, L.; Caridakis, G. Multimodal emotion recognition from expressive faces, body gestures and speech. In Proceedings of the Doctoral Consortium of ACII, Lisbon, Portugal, 12–14 September 2007.
37. Zhang, H.; Meng, Z.; Luo, M.; Han, H.; Liao, L.; Cambria, E.; Fei, H. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In Proceedings of the ACM on Web Conference 2025, Sydney, Australia, 28 April–2 May 2025; pp. 2872–2881.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.