Cureus Journal of Computer Science

Open Access Original Article

Am I as Effective at Identifying Emotions as Artificial Intelligence? A Comparative Study of Emotion Recognition

Received 06/17/2025
Review began 06/27/2025
Review ended 08/12/2025
Published 08/18/2025

© Copyright 2025
Grove et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI:
https://doi.org/10.7759/s44389-025-07224-y

Traci R. Grove [1] , Alexandra T. Lucas [2] , MaryAnn Martin [3] , Cathleen M. Deckers [4] , Lulu Sherif Mahmood [5] , Nicole Danaher-Garcia [6] , Mark W. Scerbo [7] , Suzan Kardong-Edgren [6] , Janice C. Palaganas [8, 6]

1. School of Nursing, Benedictine College, Atchison, USA 2. Division of Pediatric Critical Care, Department of Pediatrics, MassGeneral for Children, Boston, USA 3. Donna & Robert Manning College of Nursing & Health Sciences, University of Massachusetts Boston, Boston, USA 4. School of Nursing, California State University, Long Beach, USA 5. Simulation and Skill Centre, Father Muller Medical College, Mangalore, IND 6. Department of Health Professions Education, MGH Institute of Health Professions, Boston, USA 7. Psychology, Old Dominion University, Norfolk, USA 8. REBEL Lab, Center of Excellence in Healthcare Simulation Research, MGH Institute of Health Professions, Boston, USA

Corresponding author: Traci R. Grove, tgrove@benedictine.edu

Abstract

Background

Learning conversations, or dialogues aimed at deepening understanding and reflection, are deeply influenced by emotions. Effective communication is influenced by emotional intelligence - the ability to recognize, understand, and manage both one's own and others' emotions. While advances in artificial intelligence (AI) offer new tools for emotion recognition, these technologies still struggle with accurately interpreting subtle and culturally diverse emotional expressions, sparking debate about their reliability and effectiveness. This article provides a comparative analysis of human versus AI recognition of emotions during an end-of-course reflective learning conversation.

Methods

Emotions during a structured post-conference debriefing were analyzed and coded by MorphCast ("AI data") in real time and human researchers ("peer data") from a video recording of the conversation in which audio was removed. Each individual in the learning conversation also self-assessed and coded their own emotions ("self data") from the same recording. The AI and peer data were compared to the self data.

Findings

The AI dataset captured a wider range of emotions - including *anger*, *sadness*, and *disgust* - not detected by Self or Peer assessments, which predominantly identified *happy* and *neutral* emotions. *Happiness* was lower in MorphCast assessment (20.86%), higher in human assessment (59%), and highest in self-assessment (65%).

Conclusions

Human identification was more similar to self-identification of emotions during the learning conversation than AI. AI identified a broader range of emotions, especially negative ones which were largely absent in self- and peer-reports that leaned toward happy and neutral states. These differences highlight how AI, lacking contextual understanding and relying solely on facial cues, may misinterpret emotions compared to humans who integrate verbal, situational, and social information into their assessments.

Categories: AI applications, Explainable AI
Keywords: artificial intelligence, simulation, debriefing, emotions, learning conversations

Introduction

Learning conversations - dialogues aimed at deepening understanding and reflection - are an essential part of the learning process. Effective learning conversations rely on the skill of the facilitator to perceive [1-2] and interpret the emotional dynamics within a group [2-3]. The skill of being able to accurately perceive the emotional state of oneself and others during a meaningful conversation depends on a person's emotional intelligence (EI), which includes the ability to identify, understand, express and manage effective use of emotions to interact with others [4]. Emotion recognition, an important part of EI, involves recognizing both one's own and others' emotional states, which is essential for meaningful and effective communications and understanding [1].

In simulation-based education, the facilitator's recognition of the emotional dynamics of the group supports the individual and collective cognitive processes of learning that occur in the simulation process [1-3]. Learners in simulation often experience intense emotions, which can positively or negatively impact their learning outcomes [2,5-6]. While frameworks are being developed to aid facilitators in responding to learners' emotions during debriefing, they hinge upon the facilitator's ability to recognize and identify their learners' emotional states [6]. During feedback conversations (e.g., debriefing following a simulation), facilitators who have high EI are in a better position to recognize, understand, and manage emotions of their learners, but the process of identifying and interpreting emotions in real time remains challenging [7] and requires constant attention and awareness [1,4]. The facilitator must navigate ambiguous signals, be aware of personal and cultural biases, make rapid judgments, and incorporate complex contextual information, all of which can easily lead to misunderstandings or errors that can negatively impact learners [3,8]. The ability to correctly identify emotions, therefore, is critical in guiding learning conversations.

Virtual learning conversations (e.g., online conferencing) add further complexity in the process of identifying and navigating emotions. Virtual conferencing as a tool for teaching and learning has been adopted as a mainstream practice as a result of the COVID-19 pandemic. The concept of "digital emotional intelligence" is evolving to highlight the importance of developing skills to help individuals interpret emotional cues in virtual settings [9]. Research indicates that online interactions, compared to in-person, can be fraught with communication misperceptions due to the difficulty of recognizing and interpreting emotions when using a virtual platform [10].

Artificial intelligence (AI) systems have been developed to enhance human-machine interactions by detecting, classifying, and interpreting human emotional states through facial recognition [11]. AI emotion tracking programs arose in response to the desire for technology to assist in sensing, interpreting, and responding to human emotions across diverse applications - from healthcare and customer service to marketing and workplace productivity - fostering more empathetic, personalized, and effective interactions. AI could significantly enhance the facilitation of learning conversations if they are accurate. However, there is debate as to how well AI understands and can interpret human emotions. These tools first identify and track faces, then detect facial landmarks, evaluate changes in these landmarks, and then classify the data [11]. Multiple companies have developed proprietary algorithms for this process, many of which use standardized facial expression databases designed to include a diverse range of ages, ethnicities, genders, and races [7].

MorphCast Emotional Tracking [12] is AI that detects human emotions in the context of facial expression recognition [11]. MorphCast AI's emotion recognition model is trained using custom datasets including BU-4DFE - a high-resolution 4D facial expression database containing dynamic 3D sequences of six basic emotions from over 100 diverse subjects representing a wide diversity of ethnicities, ages, and genders, including Asian, Black, Hispanic/Latino, African American, and White populations [11]. Researchers suggest that AI does not take into consideration more subtle expressions and the various degrees of emotions [8]. Cultural differences in emotional expression are recognized as a broader challenge in AI empathy [13].

With the emergence of AI emotion tracking, comparative testing of human judgement versus AI analysis in the context of emotion recognition is needed prior to determining accuracy for research purposes. This study aimed to explore the accuracy of emotional analysis made by human observers, as measured by participants' self-assessed emotions during an end-of-course reflective learning conversation (or post-course debriefing) compared to AI. Specifically, we examined 1) whether each dataset (AI, Peer, Self) captured a variety of emotions; 2) the predominant emotion in each dataset; 3) whether the datasets suggested a tendency toward positive, neutral, or negative affect; and 4) the frequency of transitions between emotions.

Theoretical underpinnings of emotion recognition

This study draws upon three primary theoretical frameworks from the fields of organizational psychology, emotion science, and educational psychology: Affective Events Theory (AET) [14], Russell's Circumplex Model of Affect [15], and Brain-Based Learning Theory [16]. Together, these theories provide a foundation for understanding the variability of emotions in the learning context, the role of emotions in learning conversations, as well as the comparative effectiveness of human versus AI in recognizing those emotions.

AET, originally developed by Weiss and Cropanzano [14] for organizational behavior, is used to understand emotional dynamics in educational environments. This theory posits that any classroom event (e.g., receiving feedback, participating in discussions, or facing academic challenges) triggers emotional reactions in students and teachers. These emotional responses are influenced by individual differences (e.g., personality, mood, and prior experiences), leading to a wide range of emotions even in response to the same event. This variability is further amplified by differences in EI and the unique ways individuals interpret and process classroom events.

Russell's Circumplex Model of Affect offers a widely accepted dimensional approach to emotion classification, organizing emotional experiences along two continuous axes: valence (pleasure to displeasure) and arousal (activation to deactivation) [15]. This model enables a nuanced understanding of emotion by capturing both the quality and intensity of affective states. Its strength lies in its generalizability across cultures, making it particularly relevant when assessing the capabilities of AI systems, which may be limited in interpreting culturally specific emotional cues. The Circumplex Model serves as an analytical lens for comparing the emotional data collected by AI, human observers, and self-analyses, particularly in terms of the trends observed in positive versus negative affect and the intensity of emotional responses.

Complementing this, Brain-Based Learning Theory [16] emphasizes the critical role of emotion in learning processes. According to this theory, emotional and cognitive functions are deeply intertwined. Emotions, whether positive or negative, directly influence how learners absorb, retain, and apply information. Positive emotional states are shown to facilitate higher-order thinking and deep learning, while negative emotions - especially those that trigger stress responses like fight or flight - can inhibit cognitive functioning and block meaningful learning. This theory highlights the educator's role as a mediator of the emotional climate, suggesting that recognizing and responding to the emotional states of learners is essential for creating psychologically safe and effective learning environments.

AET acknowledges the emotional nature of learning, Russell's model [15] provides a structure for interpreting the type and intensity of emotions, while Brain-Based Learning Theory [16] underscores the educational implications of those emotional states. By integrating these conceptual frameworks, this study situates emotion recognition - whether by human educators or AI tools - as a critical component of learning conversations that has a natural variability of emotions. These theoretical frameworks were used to provide a foundation for the study's design and interpretation of results, serving as theoretical lenses- not as sources of formal, theory-driven predictions.

Materials And Methods

We used a repeated measures design using three datasets (AI, peer, and self) to compare the accuracy in the identification of emotions made by human observers versus AI, as measured by participants' self-assessed emotions and peer-assessed emotions, during an end-of-course reflective learning conversation (or post-course debriefing).

Sample

Participants were from a non-random convenience sample of post-licensure practicing providers and educators in health professions education enrolled in a healthcare simulation research course. All five participants of the course took part in this study. Participants did not know each other prior to the course.

Setting

The study took place during an in-person continuing education course on April 3, 2025, in Boston, Massachusetts.

Institutional review board

The study was deemed exempt by the MGB Institutional Review Board, as it was a planned educational session of a professional development course intended to inform the course program and curricula. Consent forms for the study, including recordings, were signed by all participants.

Instruments used

In this study, we utilized a paid version of MorphCast EmotionalTracking [12] AI, a commercial, cloud-based facial recognition and emotional detection system that embeds in video-conferencing. Dupré et al. [11] evaluated MorphCast's performance using video-based posed and spontaneous emotional expressions, reporting moderate interrater reliability among the automatic classifiers (Fleiss' Kappa = 0.47, $p < .001$). MorphCast is grounded in the widely accepted work by Ekman and Rosenberg [17] and Russell [12,15,18-20]. Zoom MorphCast overlays the Zoom Video Conferencing platform [21].

Procedures

Participants engaged in an experiential in-person course, consisting of simulation scenarios, debriefings, a focus group, and group work, all designed to promote professional development in healthcare simulation research (see Figure *1*). The day concluded with a 15-minute group reflection session aimed at creating shared learning and discussion. The structure of the reflection included a) reactions to the day, b) reflection on things that you feel are effective and the reason you feel that way, c) reflection on things that you feel need to be improved in the course and the reason you feel that way, and c) takeaways from the day. In order to employ MorphCast, video-conferencing and recording of the session were conducted using Zoom [21]. The recorded session was uploaded to Vosaic [22], a video-coding and analysis platform designed for detailed review of recorded interactions. The platform allows researchers to perform frame-by-frame coding and generate time-stamped, exportable datasets for quantitative analysis. All participants independently joined Zoom by relocating to different areas of the building, ensuring sufficient distance to prevent audio interference. The end-of-day discussion was held via Zoom for study purposes. The session provided an opportunity for both personal and group-level reflection on the emotional experiences elicited throughout the day's activities. This final video conferencing session served as the learning conversation analyzed for this study.

FIGURE 1: Study Schema

Researcher demographics

All researchers except JCP and MWS had no previous training or research experience in EI. MWS is a human factors psychologist with experience studying nonverbal behavior. JCP has been trained in and teaches courses in EI and has conducted previous studies focused on EI and expressions. Researchers were of mixed professions and gender.

Data collection

As illustrated in Figure *1*, Study Schema, MorphCast's EmotionalTracking [12] collected real-time data on participants' emotion categories and generated .csv files of each participant's emotional profile during the recorded session. A code book of 11 emotions was created in Vosaic to match the 11 emotions reported in the MorphCast generated report. The coding categories included happy, sad, surprise, angry, fear, frustrated, bored, satisfied, self-confident, neutral, and disgust. This file was uploaded to MGB Business DropBox [23] for analysis and was accessible only to the researchers. We refer to this dataset as "AI".

The Zoom video recording was downloaded from Zoom and uploaded to Vosaic, a video-coding software designed for video-based feedback and analysis [22]. Because MorphCast does not evaluate audio or verbal intonation, the audio was removed from the original recording before uploading a "no audio" version to Vosaic.

Rater training occurred by providing the researchers with a separate (non-study) video demonstration that explained the objectives of the coding activity, offered instructions on using the Vosaic software, and demonstrated how to apply each of the codes of emotion. The aim of the study was to analyze initial variability in coding; thus, we did not conduct inter-rater reliability training.

Six researchers were randomly assigned a peer to observe and code. Each researcher individually viewed the “no audio” recording of the end-of-day reflective learning conversation, simultaneously coding for emotions of their assigned peer. We refer to this as the “Peer” data.

Participants also individually viewed the recording of the end-of-day reflective learning conversation to code their own emotions retrospectively. Because this was one of the study conditions and we aimed to capture accurate emotions during the conversation, participants coded the video with audio. We refer to this as “Self” data.

Vosaic generated the .csv files used for analysis. These files were also uploaded to MGB Business Dropbox for comparative analysis.

After all coding activities were completed, the research team collectively cleaned and prepared the datasets for analysis. This entailed removing non-relevant columns from the MorphCast data, ensuring all emotion categories were correctly included, excluding low-frequency or non-relevant emotion categories, matching category labels across datasets, aligning timestamps, and generating consolidated Excel sheets with emotion percentages for comparative analysis. This data preparation was completed under the guidance of a statistician to ensure the dataset was accurate, complete, and ready for the comparative analysis.

Analysis process

Emotions during the end-of-course reflective learning conversation were analyzed by MorphCast (“AI data”), which collected data in real time during the live session, and by human researchers (“peer data”) who retrospectively coded emotions by reviewing the video recording of the same learning conversation. Each individual in the learning conversation also assessed and coded their own emotions (“self data”) from the recording. The AI and peer data were compared to the self data. The seven most frequent categories identified by Morphcast (angry, disgust, fear, happy, neutral, sad, surprise) were used for analysis. Four lower-frequency categories (bored, frustrated, satisfied, and self-confident) were excluded due to minimal occurrence across datasets, which limited their usefulness for meaningful comparative analysis.

Basic comparative analysis of coding frequencies was conducted using Excel [24]. Comparisons between conditions were analyzed using R version 4.5.0 [25]. A Friedman nonparametric test for repeated measures was used to analyze differences between the conditions, given that the same individuals were analyzed in three groupings. To calculate skewness, categories were grouped into positive emotions (happy) and negative emotions (angry, disgust, fear, sad); surprise and neutral were excluded. Skewness of the participants’ average time spent in positive versus negative categories was calculated using the skewness function from the e1071 package in R [26].

To best answer our research aim, we sought the following specific sub-analyses:

Did each dataset (AI, Peer, Self) capture a variety of emotions?

How many different emotions were captured?

Were there any emotions identified by AI and not by peer or vice versa?

What were the similarities and differences of emotional percentages?

What was the predominant emotion in each dataset (AI, Peer, Self)?

How similar or different was AI when compared to Self?

How similar or different was Peer when compared to Self?

Did any of the datasets (AI, Peer, Self) suggest a tendency in expressive variance (e.g., skew toward neutral, positive, or negative emotion)?

What is the frequency of transitions in emotions in each dataset (AI, Peer, Self)?

Sample demographics

There were five participants in the study with no previous training or research experience in EI, or formal training in acting. All five participants were post-licensure clinicians with faculty roles. The study was open to individuals of all genders; however, the final participant group consisted entirely of women. The participants were not familiar with each other prior to joining the course.

Results And Discussion

Experimental setup

This pilot study analyzed emotions displayed during a 15-minute end-of-course reflective learning conversation among five participants. Each participant's emotional expressions were recorded using Zoom and analyzed in three ways: (1) in real-time by MorphCast AI software ("AI" data), (2) by a randomly assigned peer reviewer coding the same video without audio ("Peer" data), and (3) by the participant reviewing their own video with audio and coding their emotions ("Self" data). All analyses used a common set of 11 emotion categories. The primary was to compare emotion identification across datasets.

Sub-analysis 1: Did each dataset (AI, Peer, Self) capture a variety of emotions?

1.a. How many different emotions were captured?

Self resulted in identification of three emotions: happy, neutral, and surprise. Peer evaluation captured five emotions: disgust, happy, neutral, sadness, and surprise. AI identified three emotions: happy, neutral, and surprise (Figure *2*).

1.b. Were there any emotions identified by AI and not by Peer or vice versa?

AI identified angry in 12.41% of expressions, disgust in 10%, and sadness in 16.34%, while neither Self nor Peer identified these negative emotions (Figure *2*).

1.c. What are the similarities and differences of emotional percentages?

There was a significant difference in the emotional percentages between groups in happiness (χ^2 = 7.6, df = 2, p = .02), anger (χ^2 = 10, df = 2, p = .007), disgust (χ^2 = 10, df = 2, p = .007), fear (χ^2 = 10, df = 2, p = .007), and sadness (χ^2 = 10, df = 2, p = .007) (Figure *3*). For happiness, Self and Peer reported much higher percentages than AI. The differences for anger, disgust, fear, and sadness were due to the absence of reported time for Self and Peer. In group analysis, the majority of Self and Peer emotions were identified as happy (72% in self, 59% peer) and neutral (27% self, 39% peer) (Figure *2*). AI identified neutral (24%), happy (20.86%), and sad (16.34%) as the most prevalent emotions (Figure *2*).

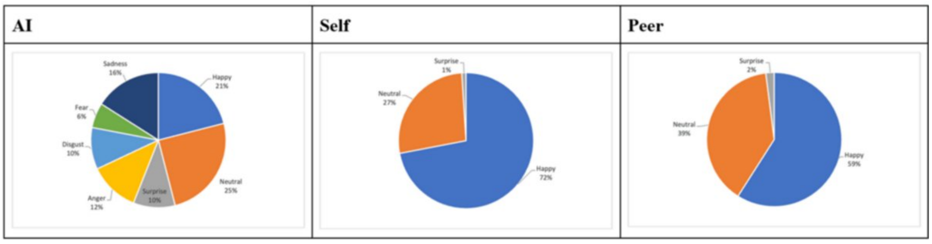

FIGURE 2: Group Percentages (Means)

Group	AI	Peer	Self	P-value
%Happiness (Median, (IQR))	10.81 (10.23,29.6)	55 (54,66)	70 (52, 83)	0.02*
%Neutral (Median, (IQR))	24.88 (18,30.7)	41 (30, 46)	30 (17, 42)	0.2
%Anger (Median, (IQR))	6.2 (3.5,11.96)	0 (0,0)	0 (0,0)	0.007*
%Disgust (Median, (IQR))	13.66 (1.7, 17.55)	0 (0,0)	0 (0,0)	0.007*
%Fear (Median, (IQR))	6.17 (2.34,9.34)	0 (0,0)	0 (0,0)	0.007*
%Sadness (Median, (IQR))	27 (19.55,27.6)	0 (0,0)	0 (0,0)	0.007*
%Surprise (Median, (IQR))	9.46 (4.96,10.1)	0 (0,4)	0 (0,0)	0.056

TABLE 1: Median Percentage of Time in Each Emotional State (Median, IQR)

An asterisk (*) next to the P-value indicates statistical significance. P-values were obtained using the Friedman test.

AI, Artificial Intelligence; IQR, Interquartile Range

Sub-analysis 2: What was the predominant emotion in each dataset (AI, Peer, Self)?

2.a. How similar or different was AI when compared to Self?

For each individual, AI identified fewer happy emotions than Self and identified between three and six additional emotions than Self (Table *1*). Sad, disgust, fear, and anger were identified in all participants by AI but not in Self (Table *1*).

2.b. How similar or different was Peer as compared to Self?

Peers generally (i.e., four of five cases) identified 0-1 additional emotions as compared to Self, and Self-identified additional emotions in one of the five cases (Table *1*). The overall percentage of time in an emotional state was similar between Self and Peer (65% vs. 59% happy and 32% vs. 39% neutral, respectively, compared to AI at 20.86% happy and 24.13% neutral) (Figure *2*).

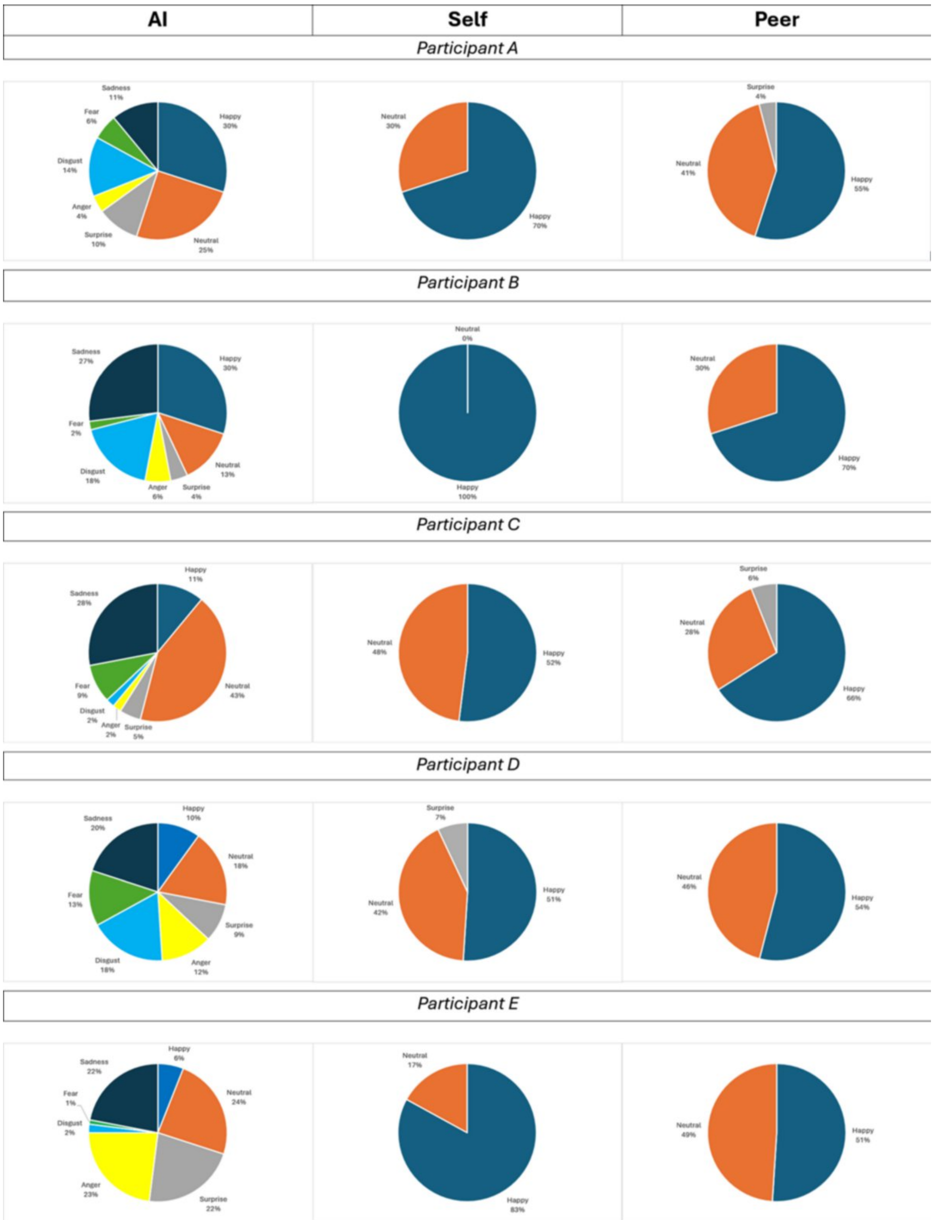

FIGURE 3: Individual Percentages (Mean)

The specific emotional breakdown for each participant across datasets is detailed in Figure 3. This illustrates how AI identified more varied emotional states compared to Peer and Self, which primarily reported happiness. In all five cases, AI detected emotions not recognized by the participant or their assigned peer reviewer.

Sub-analysis 3: Did any of the datasets (AI, Peer, Self) suggest a tendency in expressive variance (i.e., skew toward positive or negative emotion)?

AI analysis skewed towards negative emotions, with anger, disgust, fear, and sadness comprising 45% of group emotion with a strong negative skewness (skew = -32.7), compared to 1% in Self and none in Peer (Figure 4). Positive emotions (happiness) composed only 20% of emotional states in AI analysis vs. 65% self and 59% peer (Figure 4). Self and Peer demonstrated strong positive skewness (71 and 59, respectively).

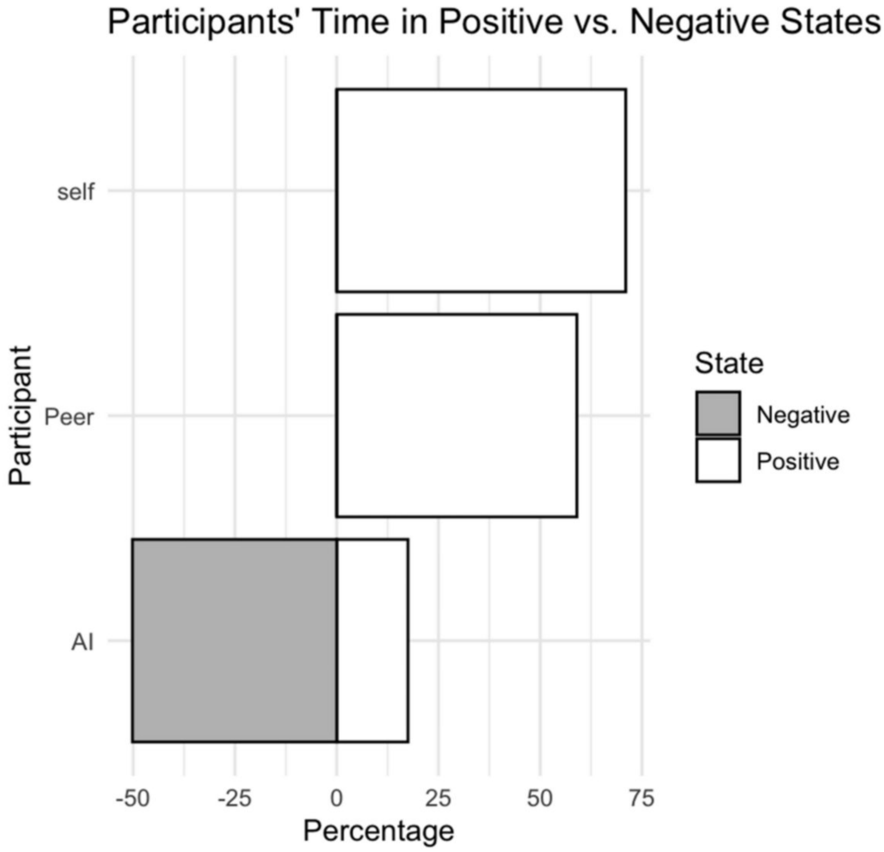

FIGURE 4: Skewness Toward Positive or Negative Emotional States by Participants' Time Per Group

The skewness of emotional distribution is illustrated in Figure 4, which clearly shows a strong negative skew in the AI dataset and a positive skew in the Peer and Self datasets. This visualization supports quantitative findings that AI disproportionately identified negative states (e.g., sadness, disgust, anger), whereas human-coded datasets leaned toward positive emotional affect, primarily happiness.

Sub-analysis 4: What is the frequency of transitions in each dataset (AI, Peer, Self)?

Transitions, defined as timepoints where an individual was observed to change from one emotion to another, were highest in peers (mean = 16±7), followed by AI (mean = 12±2.2), and lowest in self-observations (1±0.7) (Figure 5).

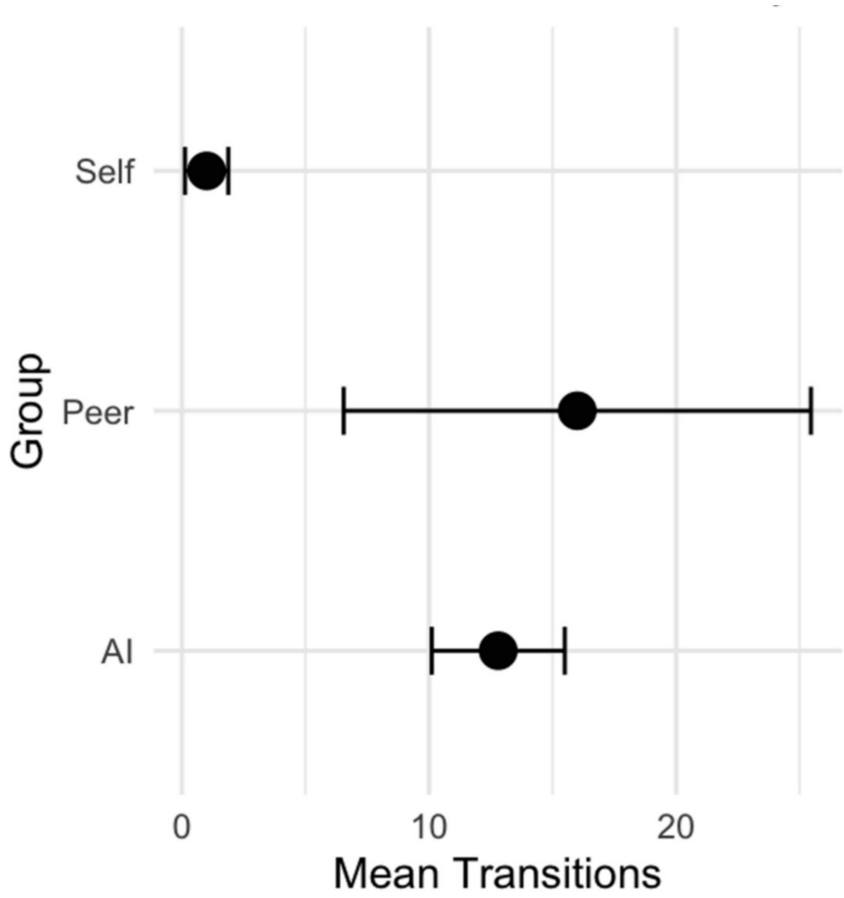

FIGURE 5: Forest Plot of Coding Transitions (Mean, Standard Deviation)

Figure 5 highlights peer observers may over-interpret subtle emotional shifts, whereas participants self-reported relatively stable emotional states.

Discussion

This study compared how AI, human observers, and individuals themselves recognized emotions during an end-of-course reflective learning conversation. The analysis showed that AI detected a wider range of emotions, including more negative ones like anger, disgust, and sadness, which were largely absent in the peer and self-analyses which skewed toward positive and neutral emotions. While peer- and self-assessed data aligned closely, AI results differed significantly ($p = .05$ in six out of seven emotional states), showing a strong bias toward negative emotional states and identified emotions not perceived by humans.

AI versus human interpretation

AI analysis differed notably from the researchers' interpretation of facial expressions, particularly in its ability to categorize a broader range of emotions. According to Ballesteros et al. [27], emotions such as anger, disgust, fear, and sadness are inherently more difficult to identify, as highlighted in Russell's Circumplex Model [15,28]. Similarly, Solis-Arrazola et al. [29] found that emotions like sad, neutral, and angry pose challenges for both AI and human detection. However, in this study, AI demonstrated a heightened sensitivity to subtle emotional cues, identifying emotions, such as disgust, that were not recognized by humans. This discrepancy may be attributed to the difficulty that humans have in detecting microexpressions, or brief, involuntary fleeting facial expressions that last less than half a second but reveal genuine emotions [30]. It may also be attributed to the human's lack of formal training in body language, including microexpressions, and emotion recognition. Research consistently demonstrates the critical importance of context in emotion perception and processing [27,31]. Studies show that contextual information, including environment, voices, and cultural/organizational factors [32], influence how emotions are perceived. Ignoring context can lead to inaccurate emotion recognition [33]. Furthermore, individual differences in environmental sensitivity, such as sensory processing, affect how people respond to positive and negative contexts [34].

While AI detected a higher frequency of negative emotions compared to human observers, human observers identified significantly more expressions of happiness than the AI did, both in peer and self-analysis. Research indicates that humans have a critical need to be seen as happy and to report happiness [35], influenced by social biases [36-37]. People tend to overestimate positive events and underestimate negative ones, exhibiting an optimism bias [35]. Social desirability bias leads individuals to portray themselves more positively than objectively observed [36]. Self-enhancement motivation distorts self-reports of subjective well-being, resulting in more positive assessments [37].

Human coding alignment despite audio context

Unlike the AI- and peer-analysis, the participants reviewed recordings that included both visual and verbal cues, allowing them to interpret emotions using facial expressions, tone of voice, and contextual information. This difference allowed participants to shape their coding decisions by additional sensory and contextual input, while AI- and peer-analysis were constrained to facial interpretation alone. Holland et al. [38] evaluated facial mimicry - or the ability to portray emotions for the purpose of emotional recognition - and stipulated that mimicry occurs as a result of creating an opportunity for social bonding and affiliation with the group demonstrating positive versus negative emotions. Interestingly, Holland et al. found that a significant correlation of facial mimicry was evident among females [38]. The participants in this comparative analysis were all female.

Emotion

Our findings suggest that human observers better approximated self-reported expressions in identifying spontaneous expressions than AI, underscoring the need to further validate and refine AI tools for real-world emotion detection. This is similar to prior studies, which have shown that AI performs significantly better at identifying posed expressions with sensitivity and precision dropping as low as 27-48% for spontaneous emotions [11,39]. Gaya-Morey et al. [40] found that humans and AI tend to focus on different facial areas when identifying emotions, raising concerns about how AI is “learning” to interpret affective cues.

Biases in human versus AI emotion interpretation are still not well understood. Jindal and Singh [41] suggest that multimodal AI that combines visual, auditory, and other sensory inputs is more accurate than single-modality systems in detecting microexpressions and subtle emotional shifts. Additional factors like lighting, camera angle, distance from the subject, and environmental context also influence AI accuracy [12]. In our study, we did not evaluate these variables and so cannot control for these effects. Lomas et al. [42] emphasize that emotional expression can vary depending on the AI used and the specific emotions being analyzed. Jindal and Singh [41] further note the importance of vocal tone, speech patterns, and body language in emotion recognition. The use of multimodal AI may eliminate the level of biases seen in AI. Future studies may better evaluate the impact of these factors.

Limitations

This study has several limitations. First, the design was restricted by the available tools. For best comparative analysis, we were limited to the MorphCast-coded emotion categories, and only later realized the system offered an extended range of emotions beyond those originally coded. Second, AI dynamically assessed facial expressions on a minute-by-minute basis and human coders were less likely to apply emotion codes unless there was a noticeable or obvious change in facial expression, introducing a potential underrepresentation of microexpressions. Third, self-assessment may carry retrospective bias where participants' memories and interpretations of their emotions may not always align with their actual in-the-moment expressions, creating inconsistency between the self-reported data and the observed data [43]. The self-analysis was coded with audio, providing context that may have changed the analysis of peer and AI. Additionally, the peer-analysis was initially coded with audio in March, then recoded in May without audio when we discovered that MorphCast only coded using facial recognition. This potentially influenced the way emotions were perceived and coded. Finally, participants were not prompted to code at any certain interval, which might have ensured that humans coded observations consistently, improving the consistency and comparability of the data across participants and to AI. The lack thereof potentially affected the accuracy and completeness of coded data.

While the study involved practicing clinicians and higher education faculty, the small nonrandom convenience sample, single virtual reflective learning conversation, and all-female composition provide only suggestive evidence of the difference between human and AI interpretation of emotions. While no prior personal or professional relationships existed between the participants and the researchers, all were taking part in an in-person course together, which may introduce bias.

A key limitation in any emotion recognition study is that outward expressions do not always reflect an individual’s true emotional state. People often mask their genuine feelings due to social norms, cultural expectations, or emotional self-protection [11], which can lead to misinterpretation by both AI and human observers. As a result, relying solely on facial expressions or observable cues may oversimplify the complexity of actual emotional experiences.

Future directions

Given the rapid advancement of AI capabilities and the critical role of emotions in learning conversations, continued research is both urgent and essential to ensure responsible development and application, as well as effective technology-assisted education. This may include manipulation of AI through standardized interventions during the conversation for specific reactions and accuracy testing. The identification of use cases may also inform AI developers and industry partners. For example, future applications might include wearable AI devices, such as smart glasses equipped with emotion recognition technology to provide real-time emotional feedback during conversations. Opportunities for future research in this area are vast. Further exploration of the integration of multimodal AI that incorporate auditory, contextual interpretation, and verbal cues alongside facial recognition is needed to more accurately reflect how humans interpret emotions in real-world interactions. Expanding the dataset to include larger and more diverse participant groups (i.e., gender, age, and culture) could improve generalizability and help identify patterns in emotion recognition variance. Training human coders in recognizing microexpressions and providing structured prompts during analysis could reduce subjective bias and improve coding consistency. Additional research should also investigate how contextual understanding influences emotional interpretation, especially in human participants who may rely more heavily on tone, language, and situational cues. Finally, future studies should examine the ethical implications of AI's heightened sensitivity to negative emotions, particularly in educational and clinical settings, and assess how AI-detected emotional data can be responsibly used without reinforcing misinterpretation or bias.

Conclusions

While AI like MorphCast shows promise in emotion recognition, they currently lack the nuance and contextual understanding that human observers bring. Further research and more advanced training models are essential to improve AI's ability to accurately interpret both posed and spontaneous emotional expressions. Until then, human interpretation remains the more reliable approach in emotionally complex learning environments.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Acquisition, analysis, or interpretation of data: Traci R. Grove, Alexandra T. Lucas, MaryAnn Martin, Cathleen M. Deckers, Lulu Sherif Mahmood, Nicole Danaher-Garcia, Suzan Kardong-Edgren, Janice C. Palaganas, Mark W. Scerbo

Drafting of the manuscript: Traci R. Grove, Alexandra T. Lucas, MaryAnn Martin, Cathleen M. Deckers, Lulu Sherif Mahmood, Nicole Danaher-Garcia, Suzan Kardong-Edgren, Janice C. Palaganas, Mark W. Scerbo

Critical review of the manuscript for important intellectual content: Traci R. Grove, Alexandra T. Lucas, MaryAnn Martin, Cathleen M. Deckers, Lulu Sherif Mahmood, Nicole Danaher-Garcia, Suzan Kardong-Edgren, Janice C. Palaganas, Mark W. Scerbo

Supervision: Traci R. Grove, Janice C. Palaganas

Concept and design: Janice C. Palaganas

Disclosures

Human subjects: Consent was obtained or waived by all participants in this study. Mass General Brigham Institutional Review Board issued approval 025P001019. This study was reviewed by the Mass General Brigham Institutional Review Board and was deemed exempt. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

This research was done collaboratively by the faculty and Cohort 1 of The Healthcare Simulation Research Course at MGH Institute of Health Professions. The authors would like to thank Drs. Jared Kutzin and Debra Nestel for their thoughts on the development of this study.

References

1. Boden M, Thompson R: Meta-analysis of the association between emotional clarity and attention to emotions. Emotion Review. 2016, 9:79-85. 10.1177/1754073915610640
2. George J: Emotions and leadership: The role of emotional intelligence. Human Relations. 2000, 53:1027-1055. 10.1177/0018726700538001
3. Šimić G, Tkalčić M, Vukić V, et al.: Understanding emotions: Origins and roles of the amygdala. Biomolecules. 2021, 11:823. 10.3390/biom11060823
4. Kotsou I, Mikolajczak M, Heeren A, Grégoire J, Leys C: Improving emotional intelligence: A systematic review of existing work and future challenges. Emotion Review. 2018, 11:151-165. 10.1177/1754073917735902
5. Ahn B, Maurice-Ventouris M, Bilgic E, et al.: A scoping review of emotions and related constructs in simulation-based education research articles. Advances in Simulation. 2023, 8:22. 10.1186/s41077-023-00258-z
6. LeBlanc VR, Brazil V, Posner GD: More than a feeling: emotional regulation strategies for simulation-based education. Advances in Simulation. 2024, 9:53. 10.1186/s41077-024-00325-z
7. Krumhuber EG, Küster D, Namba S, Skora L: Human and machine validation of 14 databases of dynamic facial expressions. Behavior Research Methods. 2021, 53:686-701. 10.3758/s13428-020-01443-y
8. Parada-Cabaleiro E, Batliner A, Schmitt M, Schedl M, Costantini G, Schuller B: Perception and classification of emotions in nonsense speech: Humans versus machines. PLoS ONE. 2023, 18:e0281079. 10.1371/journal.pone.0281079
9. Audrin C, Audrin B: More than just emotional intelligence online: Introducing "digital emotional intelligence". Frontiers in Psychology. 2023, 14:1154355. 10.3389/fpsyg.2023.1154355
10. Lin T, Anderson T: Reduced therapeutic skill in teletherapy versus in-person therapy: The role of non-verbal communication. Counselling and Psychotherapy Research. 2023, 24:317-327. 10.1002/capr.12666
11. Dupré D, Krumhuber EG, Küster D, McKeown GJ: A performance comparison of eight commercially available automatic classifiers for facial affect recognition. PLoS ONE. 2020, 15:e0231968. 10.1371/journal.pone.0231968
12. MorphCast: MorphCast Emotion AI. Accessed: April 3, 2025: https://morphcast.com.
13. Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D: Capacity of generative AI to interpret human emotions from visual and textual data: Pilot evaluation study. JMIR Mental Health. 2024, 11:e54369. 10.2196/54369
14. Weiss HM, Cropanzano R: Affective Events Theory: A theoretical discussion of the structure, causes and consequences of affective experiences at work. Research in Organizational Behavior. 1996, 18:1-74.
15. Russell JA: A circumplex model of affect. Journal of Personality and Social Psychology. 1980, 39:1161-1178. 10.1037/h0077714
16. Caine G, Caine RN: Meaningful learning and the executive functions of the brain. New Directions for Adult & Continuing Education. 2006, 2006:53-61. 10.1002/ace.219
17. Ekman P, Rosenberg EL: What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS) (2nd edn). Oxford University Press, New York; 2005. 10.1093/acprof:oso/9780195179644.001.0001
18. Ekman P: Emotions revealed. BMJ. 2004, 328:0405184. 10.1136/sbmj.0405184
19. Ekman P, Friesen WV, Ellsworth P: Emotion in the Human Face: Guidelines for Research and an Integration of Findings. Elsevier, New York; 2013.
20. Sreeja PS, Mahalakshmi GS: Emotion models: A review. International Journal of Control Theory and Applications. 2017, 10:651-657.
21. Zoom Video Communications, Inc.: Zoom (Version 5.0). Accessed: April 3, 2025: https://zoom.us.
22. Vosaic: Vosaic Video Analysis. Accessed: April 3, 2025: https://vosaic.com.
23. Dropbox Inc.: Dropbox (Version 219.4.4463). Accessed: April 3, 2025: https://www.dropbox.com/features.
24. Excel: Microsoft Excel (Version 16.0). Accessed: April 3, 2025: https://office.microsoft.com/excel.
25. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna; 2025.
26. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F: e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. (2024). https://cran.r-project.org/web/packages/e1071/index.html.
27. Ballesteros J, Ramirez G, Moreira F, Solano A, Pelaez C: Facial emotion recognition through artificial intelligence. Frontiers in Computer Science. 2024, 6:1359471. 10.3389/fcomp.2024.1359471
28. Russell JA, Lewicka M, Niit T: A cross-cultural study of a circumplex model of affect. Journal of Personality and Social Psychology. 1989, 57:848-856. 10.1037/0022-3514.57.5.848
29. Solis-Arrazola M, Sanchez-Yanez R, Gonzalez-Acosta A, Garcia-Capulin C, Rostro-Gonzalez H: Eliciting emotions: Investigating the use of generative AI and facial muscle activation in children's emotional recognition. Big Data and Cognitive Computing. 2025, 9:15. 10.3390/bdcc9010015
30. Adegun I, Vadapalli H: Facial micro-expression recognition: A machine learning approach. Scientific African. 2020, 8:e00465. 10.1016/j.sciaf.2020.e00465
31. Hess U, Kafetsios K: Infusing context into emotion perception impacts emotion decoding accuracy: A truth and bias model. Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie). 2021, 68:285-294. 10.1027/1618-3169/a000531
32. Barrett LF, Mesquita B, Gendron M: Context in emotion perception. Current Directions in Psychological Science. 2011, 20:286-290. 10.1177/0963721411422522
33. Marpaung A, Gonzalez A: Can an affect-sensitive system afford to be context independent?. Modeling and Using Context. Brézillon P, Turner R, Penco C (ed): Springer, Cham; 2017. 10257:454-467. 10.1007/978-3-319-57837-8_38
34. Greven CU, Homberg JR: Sensory processing sensitivity—For better or for worse? Theory, evidence, and

societal implications. The Highly Sensitive Brain: Research, Assessment, and Treatment of Sensory Processing Sensitivity. Acevedo BP (ed): Academic Press, Cambridge; 2020. 51-74. 10.1016/B978-0-12-818251-2.00003-5

35. Chang EC: Optimism & Pessimism: Implications for Theory, Research, and Practice. American Psychological Association, Washington, DC; 2000. 10.1037/10385-000
36. Beins BC: Social desirability bias. The Encyclopedia of Cross-Cultural Psychology. Keith KD (ed): Wiley, Malden, MA; 2013. 10.1002/9781118339893.wbeccp501
37. Wojcik SP, Ditto PH: Motivated happiness. Social Psychological and Personality Science. 2014, 5:825-834. 10.1177/1948550614534699
38. Holland AC, O'Connell G, Dziobek I: Facial mimicry, empathy, and emotion recognition: a meta-analysis of correlations. Cognition & Emotion. 2020, 35:150-168. 10.1080/02699931.2020.1815655
39. Kuntzler T, Höfling TTA, Alpers GW: Automatic facial expression recognition in standardized and non-standardized emotional expressions. Frontiers in Psychology. 2021, 12:627561. 10.3389/fpsyg.2021.627561
40. Gaya-Morey FX, Ramis-Guarinos S, Manresa-Yee C, Buades-Rubio JM: Unveiling the human-like similarities of automatic facial expression recognition: An empirical exploration through explainable AI. Multimedia Tools and Applications. 2024, 83:85725-85753. 10.1007/s11042-024-20090-5
41. Jindal V, Singh K: Emotion recognition in AI: Bridging human expressions and machine learning. International Journal For Multidisciplinary Research. 2025, 7:1-10. 10.36948/ijfmr.2025.v07i01.34795
42. Lomas JD, van der Maden W, Bandyopadhyay S, et al.: Evaluating the alignment of AI with human emotions. Advanced Design Research. 2024, 2:88-97. 10.1016/j.ijadr.2024.10.002
43. Motavalli A, Nestel D: Complexity in simulation-based education: exploring the role of hindsight bias. Advances in Simulation. 2016, 1:3. 10.1186/s41077-015-0005-7