frontiers | Research Topics

Emotions and artificial intelligence

Edited by
Simone Belli, Angel Barrasa, Cristian López Raventós
and Mariacarla Martí-González

Coordinated by
Miriam Jiménez Bernal

Published in
Frontiers in Psychology
Frontiers in Artificial Intelligence
Frontiers in Computer Science

ISSN 1664-8714
ISBN 978-2-8325-7327-3
DOI 10.3389/978-2-8325-7327-3

Generative AI statement
Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Emotions and artificial intelligence

Topic editors

Simone Belli — Complutense University of Madrid, Spain
Angel Barrasa — University of Zaragoza, Spain
Cristian López Raventós — National Autonomous University of Mexico, Mexico
Mariacarla Martí-González — Complutense University of Madrid, Spain

Topic coordinator

Miriam Jiménez Bernal — European University of Madrid, Spain

Citation

Belli, S., Barrasa, A., López Raventós, C., Martí-González, M., Bernal, M. J., eds. (2026). *Emotions and artificial intelligence*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-7327-3

Table of contents

frontiers | Frontiers in Psychology

TYPE Editorial
PUBLISHED 15 December 2025
DOI 10.3389/fpsyg.2025.1720713

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Wataru Sato,
RIKEN, Japan

*CORRESPONDENCE
Simone Belli
✉ sbelli@ucm.es

RECEIVED 08 October 2025
ACCEPTED 24 November 2025
PUBLISHED 15 December 2025

CITATION
Belli S, Barrasa A, López Raventós C and Marti M (2025) Editorial: Emotions and artificial intelligence. *Front. Psychol.* 16:1720713. doi: 10.3389/fpsyg.2025.1720713

Editorial: Emotions and artificial intelligence

Simone Belli[1,2]*, Angel Barrasa[3], Cristian López Raventós[4] and Mariacarla Marti[1]

[1]Complutense University of Madrid, Madrid, Spain, [2]Universidad Ecotec, Samborondon, Ecuador, [3]Universidad de Zaragoza, Zaragoza, Spain, [4]Universidad Nacional Autonoma de Mexico Escuela Nacional de Estudios Superiores Unidad Morelia, Morelia, Mexico

KEYWORDS

emotions, artificial intelligence, social psychology, ChatGPT, music

Editorial on the Research Topic
Emotions and artificial intelligence

In 2023, we launched this Research Topic for *Frontiers in Psychology* to explore the intersection between emotions and Artificial Intelligence (AI). That year marked a turning point, the impact of generative AI expanded rapidly, influencing nearly every aspect of our lives. This profound transformation prompted us to open a dedicated space for reflection and scientific dialogue. At the time, there was a notable gap in the scientific literature regarding the psychological dimensions of AI, particularly how emotions are shaped by, interpreted through, or integrated into artificial systems. Our aim was to address this gap and encourage interdisciplinary contributions bridging psychology, emotion studies, and AI technologies.

Over the course of this initiative, we received numerous valuable submissions. After careful review, we selected 15 contributions that we believe best represent the current advancements and perspectives in this emerging field. These works significantly enhance our understanding of the emotional dimensions of human-AI interaction, and we are proud to present them in this Research Topic.

In the paper titled "*Social and ethical impact of emotional AI advancement: the rise of pseudo-intimacy relationships and challenges in human interactions*" (Wu), author argues that the integration of emotional intelligence into algorithmic platforms is ushering in a new human interaction model: the pseudo-intimacy relationship. The goal of the work is to theoretically define this pseudo-intimacy and conclude that, despite the tensions and contradictions EAI introduces into the social environment, its technological progress can and should continue, provided that its profound impact on existing social paradigms and its ethical challenges are fully addressed.

The paper, "*Intrinsic motivation in cognitive architecture: intellectual curiosity originated from pattern discovery*" (Nagashima et al.), proposes a novel mechanism for intrinsic motivation rooted in the perspective of human cognition, specifically defining intellectual curiosity as the drive to discover novel, compressible patterns in data. Through simulations involving three ACT-R models with varying levels of thinking navigating different mazes, the study found that increasing intellectual curiosity negatively impacted task completion in models with lower thinking levels but positively affected those with higher thinking levels.

The article, "*The role of socio-emotional attributes in enhancing human-AI collaboration*" (Kolomaznik et al.), investigates how incorporating socio-emotional attributes like trust, empathy, and rapport can significantly optimize human-AI interactions and boost productivity. The analysis suggests that when AI is designed to align with human emotional and cognitive needs, it fosters deeper trust and empathetic understanding, leading to marked improvements in collaborative efficiency, productivity, and the ethical integrity of human-AI relationships.

The exploratory study, "*Emotion topology: extracting fundamental components of emotions from text using word embeddings*" (Plisiecki and Sobieszek), investigates the potential of word embeddings as a novel, data-driven method for emotion decomposition analysis. The study concludes that word embeddings are a promising, theory-agnostic tool for uncovering emotional nuances and suggests that this methodology could be broadly applied to enrich the understanding of emotional and other psychological constructs in an ecologically valid way.

The article "*Emotional responses of Korean and Chinese women to Hangul phonemes to the gender of an artificial intelligence voice*" (Lee et al.) investigates how cultural background and AI voice gender influence emotional responses to phonemes. The findings demonstrate that even phonemic units without semantic meaning can elicit varying emotional responses depending on both cultural context and AI voice gender.

The article "*Impact of media dependence: how emotional interactions between users and chat robots affect human socialization?*" (Yuan et al.) explores how media dependence shapes emotional interactions between users and chatbots, focusing on the Replika platform. Results indicate that while most users are light users who invest moderate time and emotions, chatbot interactions provide satisfaction, companionship, and relief from loneliness.

The article "*Is it possible for people to develop a sense of empathy toward humanoid robots and establish meaningful relationships with them?*" (Morgante et al.) presents a systematic review on empathy in human–robot interaction (HRI). Findings suggest that empathy is more likely when robots display anthropomorphic traits, such as facial expressions, gestures, or emotional narratives. The authors conclude that further research is needed to refine empathy models in robotics while ensuring responsible and beneficial applications for society.

The article "*A research on copyright issues impacting artists emotional states in the framework of artificial intelligence*" (Kambur and Dolunay) examines how copyright issues, particularly in the context of AI-generated art, impact artists' emotional states and creative motivation. The research emphasizes the importance of updating national and international copyright laws to address digital and AI-related works to better protect artists' rights and emotional wellbeing. The study advocates for increased awareness and educational activities to raise understanding of copyright issues.

In "*Implementing machine learning techniques for continuous emotion prediction from uniformly segmented voice recordings*," Diemerling et al. introduce an innovative approach using neural networks to detect emotions in short voice segments in real time, demonstrating promising results for enhancing human-machine interaction.

Dolunay and Temel, in "*The relationship between personal and professional goals and emotional state in academia*," examine how academics' aspirations and emotions influence the ethical use of AI, emphasizing the need for training programs and institutional ethics to prevent unethical conduct.

Tretter, in "*Equipping AI-decision-support-systems with emotional capabilities? Ethical perspectives*," discusses the ethical implications of endowing decision-support systems with emotional capacities, advocating for a balance between potential benefits and risks of manipulation and accountability loss.

Vistorte et al., in "*Integrating artificial intelligence to assess emotions in learning environments*," conduct a systematic review of the current state of AI in emotional assessment within education, highlighting its potential to personalize learning but warning of challenges related to accuracy, privacy, and ethics. Together, these studies demonstrate the importance of advancing the use of emotional AI responsibly, combining technological innovation with deep ethical and social reflection.

The study, "*Decoding emotional responses to AI-generated architectural imagery*" by Zhang et al., investigated how AI-generated images evoke emotion and whether an architectural background influences that perception. The results showed that AI is effective at conveying positive emotions, particularly joy in interior settings, but struggles with negative ones.

The study "*Artificial intelligence and social intelligence: preliminary comparison study between AI models and psychologists*" by Sufyan et al. examined the social intelligence (SI) of three AI models against that of 180 psychology students and doctoral candidates. The study concludes that AI's ability to understand emotions and social behavior is developing at a rapid pace and suggests that these models could become valuable tools in counseling and psychotherapy.

A study by Kumar and Kathiravan titled "*Emotion recognition from MIDI musical file using Enhanced Residual Gated Recurrent Unit architecture*" investigates a new method for detecting emotions in music. The researchers used a sophisticated AI model called the Enhanced Residual Gated Recurrent Unit (RGRU) to analyze MIDI music files. The findings highlight the potential for using this technology to create more advanced music recommendation systems.

The 15 contributions gathered in this Research Topic successfully addressed the initial scientific gap, offering a rich, interdisciplinary perspective on the intersection of emotions and Artificial Intelligence at a pivotal time of rapid generative AI expansion. Collectively, these works reveal a dual imperative: to foster technological advancement in emotional AI while rigorously upholding ethical and human-centric principles. The Research Topic establishes that AI is fundamentally reshaping psychological and social paradigms. For instance, the concept of the pseudo-intimacy relationship highlights the need to understand how human emotional needs are being met, and potentially substituted, by AI, urging developers and policymakers to address risks like human alienation proactively. Simultaneously, research into intrinsic motivation shows that sophisticated cognitive architectures can model complex human drives, suggesting that future AI innovation is closely tied to designing systems that genuinely emulate higher-level thinking, curiosity, and goal-directed intention. Furthermore, the findings on the enhanced social intelligence of advanced AI models compared to human psychologists underscore AI's rapid development in simulating human social cognition, pointing to its growing role as a potential tool in fields like counseling and psychotherapy. A central theme across the articles is the necessity of a holistic, human-centric approach to AI development. Studies consistently advocate for incorporating socio-emotional attributes like trust and empathy into AI design to optimize collaboration and productivity. However, this progress is tempered by ethical warnings: the review on empathy in human-robot interaction raises concerns about emotional manipulation, while papers on decision-support systems and academic ethics stress the dangers of accountability loss and unethical conduct. Crucially, the research on copyright issues underscores that legal and institutional frameworks must adapt to protect artists' emotional wellbeing and creative motivation in the face of AI-generated content. These findings collectively demonstrate that advancing emotional AI responsibly requires balancing technological innovation with robust ethical reflection and public literacy. Moreover, the research significantly advances the methodology for understanding and measuring emotion; the validation of word embeddings for emotion decomposition analysis provides a novel, data-driven tool for psychological research. Simultaneously, studies focusing on specific applications, such as continuous emotion prediction from voice, emotion recognition in musical files, and the decoding of emotional responses to AI-generated architectural imagery, showcase the practical potential of these technologies across diverse industries. Critically, the identification of significant cross-cultural differences in emotional responses to AI voice gender emphasizes that future AI applications must prioritize cultural nuance and customization to achieve truly optimized and effective human-machine interaction. This body of work provides a solid foundation, calling for continued research to refine models of empathy and motivation, establish clearlegal frameworks, and ensure that AI's expanding emotional capabilities are deployed for the beneficial and ethical advancement of society.

Author contributions

SB: Writing – original draft, Writing – review & editing. AB: Writing – original draft, Writing – review & editing. CL: Writing – review & editing, Writing – original draft. MM: Writing – review & editing, Writing – original draft.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 21 December 2023
DOI 10.3389/fcomp.2023.1305413

OPEN ACCESS

EDITED BY
Mariacarla Martí-González,
University of Valladolid, Spain

REVIEWED BY
N. Bharathiraja,
Chitkara University, India
Jiri Pribil,
Slovak Academy of Sciences, Slovakia

*CORRESPONDENCE
M. Kathiravan
✉ mkathiravan@hindustanuniv.ac.in

RECEIVED 01 October 2023
ACCEPTED 30 November 2023
PUBLISHED 21 December 2023

CITATION
Kumar VB and Kathiravan M (2023) Emotion recognition from MIDI musical file using Enhanced Residual Gated Recurrent Unit architecture. *Front. Comput. Sci.* 5:1305413. doi: 10.3389/fcomp.2023.1305413

Emotion recognition from MIDI musical file using Enhanced Residual Gated Recurrent Unit architecture

V. Bhuvana Kumar and M. Kathiravan*

Computer Science and Engineering, Hindustan Institute of Technology and Science, Chennai, India

The complex synthesis of emotions seen in music is meticulously composed using a wide range of aural components. Given the expanding soundscape and abundance of online music resources, creating a music recommendation system is significant. The area of music file emotion recognition is particularly fascinating. The RGRU (Enhanced Residual Gated Recurrent Unit), a complex architecture, is used in our study to look at MIDI (Musical Instrument and Digital Interface) compositions for detecting emotions. This involves extracting diverse features from the MIDI dataset, encompassing harmony, rhythm, dynamics, and statistical attributes. These extracted features subsequently serve as input to our emotion recognition model for emotion detection. We use an improved RGRU version to identify emotions and the Adaptive Red Fox Algorithm (ARFA) to optimize the RGRU hyperparameters. Our suggested model offers a sophisticated classification framework that effectively divides emotional content into four separate quadrants: positive-high, positive-low, negative-high, and negative-low. The Python programming environment is used to implement our suggested approach. We use the EMOPIA dataset to compare its performance to the traditional approach and assess its effectiveness experimentally. The trial results show better performance compared to traditional methods, with higher accuracy, recall, *F*-measure, and precision.

KEYWORDS

emotion recognition, Musical Instrument Digital Interface, Enhanced Residual Gated Recurrent Unit, adaptive Red Fox algorithm, EMOPIA

1 Introduction

The essence of music is deeply intertwined with emotion, as the emotional landscape of a musical piece can shift dramatically with variations in intensity, speed, and length. According to numerous studies (Juslin and Timmers, 2010; Ferreira and Whitehead, 2021), the close connection between musical structures and emotions has received a lot of attention in recent research, particularly in the fields of affective music composition and music emotion analysis. These investigations underscore the necessity of understanding how music's structural components influence emotional expression, a critical aspect for machines to effectively communicate and interact with human emotions (Koh and Dubnov, 2021). A song's mood can elicit a wide range of emotional responses from the listener (Krumhansl, 2002), with musical conventions like scale modes, dissonance, melody motion, and rhythm consistency playing a crucial role. The belief that the fundamental structure of music is the key to eliciting emotion has increased interest in music emotion recognition (MER) research (Chen et al., 2015; Panda et al., 2018). However, this field still faces numerous challenges and unresolved issues, particularly in the identification of emotions in audio music signals. A significant hurdle in MER is the subjective nature of emotional interpretation, as individuals may experience varying emotions when listening to the same piece of music. Another challenge lies in the need for standardized, high-quality audio emotion databases. Most musical notation software supports the MIDI format, which is common in symbolic music (Hosken, 2014; Li et al., 2018) and encapsulates the messages needed to create music with electronic instruments (Good, 2001; Renz, 2002; Nienhuys and Nieuwenhuizen, 2003). Recognizing emotions in MIDI musical files is crucial for enhancing the emotional impact of music, personalizing musical experiences, enabling music therapy, and advancing our understanding of music's emotional components (Luck et al., 2008; Nanayakkara et al., 2013). However, variations in emotion recognition can occur due to the dependency on the structure and properties of the MIDI file (Bresin and Friberg, 2000; Modran et al., 2023). MIDI files mostly show technical things like tempo and musical notation. They might not have the expressive range that performance dynamics, tone, and nuance can show. Additionally, the availability and quality of labeled emotional MIDI datasets may be limited (Shou et al., 2013). To address these challenges, this study introduces several contributions. We aim to recognize and extract statistical, harmonic, rhythmic, and dynamic elements from MIDI files. We use these features to improve a recognition model that is based on a better residual gated recurrent unit architecture. This model includes an adaptive algorithm, 'Neurons', for optimizing hyper-parameters like learning rate and GRU count. The proposed paradigm categorizes emotions into four quadrants: positive-high, positive-low, negative-high, and negative-low. It was implemented on the Python platform and evaluated using the EMOPIA dataset. The effectiveness of this approach is assessed using metrics such as accuracy, *F*-measure, precision, and recall.

2 Related works

The work at Panda et al. (2018) suggested adding different audio elements that are emotionally significant to fix the problems with current technology and get around their limitations. The researchers analyzed established frameworks and categorized their often-used audio elements into eight distinct musical groupings. A public dataset of 900 audio samples with subjective comments organized according to Russell's emotion quadrants was generated to assess their research efforts. Twenty cycles of 10-fold cross-validation were used to test the audio features that were already there (baseline) and the new features that were suggested for the novel. The *F*1-score was a noticeable 9% higher, or 76.4% higher, than the *F*1-score that was obtained using the proposed features along with the same number of baseline-only characteristics. The methodology has limitations in properly detecting alterations in emotional states. The paper Bhatti et al. (2016) advocated the utilization of brain signals as a means to discern human emotions in reaction to audio music tracks. The methodology utilized the readily accessible Narosky E.E.G. equipment to capture electroencephalogram (EEG) waves. In a controlled environment, individuals were instructed to engage in passive listening to audio recordings of music, with each genre lasting for 1 min. The main objectives of this study were to ascertain the age cohorts that exhibited greater receptivity to music and to assess the influence of different musical genres on human emotions. To accurately identify human emotions, the classifier included characteristics derived from three distinct domains: time, frequency, and wavelet. These features were retrieved from recorded EEG data. The study's findings unequivocally demonstrated that utilizing a multi-layer perceptron (MLP) model, incorporating a fusion of brain signal characteristics yielded remarkably high levels of accuracy in discerning human emotional states in response to audio-music stimuli. The article Hsu et al. (2017) introduced a computerized system that utilizes electrocardiogram (ECG) data to detect and classify human emotions. Firstly, the authors employ a musical induction technique to elicit the genuine emotional states of individuals and collect their ECG signals in a non-controlled laboratory setting. Subsequently, an algorithm was developed to enable automated detection of emotions by analyzing ECG signals, specifically targeting the emotional responses evoked in individuals through music perception. Using time-, frequency-, and non-linear methods to extract physiological ECG features allowed for the identification of emotion-relevant components and their correlation with emotional states. After that, a sequential forward floating selection-kernel-based class separability-based (SFFS-KBCS-based) feature selection algorithm is created to effectively find important ECG features connected to emotions and reduce the size of the chosen features. Furthermore, generalized discriminant analysis (GDA) is employed in this process. The research work Ghatas et al. (2022) introduced a method for automating piano difficulty estimation in symbolic music using deep neural networks. The researchers employ a computational model to replicate a piano recital based on a symbolic music MIDI file. Furthermore, the components of the piano roll were disassembled. Ultimately, a model was trained to utilize components assigned to difficulty labels. Our models were evaluated using both full-track and partial-track difficulty classification problems. Numerous deep convolutional neural networks have been both theorized and empirically examined. Combined with manually crafted features, the proposed hybrid deep model demonstrated exceptional performance, achieving a state-of-the-art *F*1 score of 76.26%. This achievement represents a significant improvement, with a relative *F*1 score gain of over 10% compared to previous studies. In their publication, Hung et al. (2021) introduced a novel public dataset called EMOPIA, which consists of a medium-scale collection of pop piano recordings accompanied by emotion descriptors. The given dataset includes a variety of types of data, such as MIDI transcriptions of compositions that only use piano, as well as emotional annotations at the clip level that are organized into four separate groups. The authors were provided with prototypes of models for categorizing musical emotions at the clip level and generating symbolic music based on emotions. These models were trained on the dataset and employed state-of-the-art techniques for their respective tasks. The findings indicated that the transformer-based model demonstrated a limited ability to generate music that elicited a predetermined emotional response. The researchers accurately categorized emotions in both four quadrant and valence-wise classifications. The work Ma et al. (2022)

proposes a music creation model that incorporates emotional aspects and structural elements to make music. For making music, the suggested method used a conditional auto-regressive generative Gated Recurrent Unit (GRU) model. The authors collaborate to collectively optimize a perceptual loss and a cross-entropy loss throughout the training procedure. This optimization aims to enhance the emotional expression of the generated MIDI samples, closely resembling the original samples' emotional qualities. The results of both subjective and objective tests show that this model can create emotionally moving musical pieces that are very close to the structures that were given. Nevertheless, the system must build a comprehensive framework for evaluating the emotional impact of music. In the study Abboud and Tekli (2020), we introduced MUSEC, an innovative algorithmic framework designed for autonomous music sentiment-based expression and composition. The system identified six primary human emotions expressed in MIDI musical files: anger, fear, joy, love, sadness, and surprise. Subsequently, it generated novel polyphonic (pseudo) thematic compositions that properly conveyed the emotions above. The study's primary objective was to create a music composer grounded in sentimentality. The effectiveness of MUSEC was assessed in terms of feature parsing, sentiment expression, and music composition time. The technique has shown promise across various domains, such as music information retrieval, music composition, aided music therapy, and emotional intelligence. The research Malik et al. (2017) suggested a way to use layered convolution and recurrent neural networks to continuously predict emotions in the V-A space, which is only two dimensions. After setting up a single convolutional neural network (CNN) layer, the researchers used two separate types of recurrent neural networks (RNNs). These had each been trained in a different way to deal with arousal and valence. The methodology was evaluated using the "Media Eval 2015 Emotion in Music" dataset. To test how well the proposed Convolutional Recurrent Neural Network (CRNN) worked, sequences of different lengths were used. The results indicated that shorter durations exhibited superior performance compared to longer durations. The CRNN model shows that it can get information similar to baseline features by only using Mel-band features. Log Mel-band energy characteristics are suggested as a substitute for the baseline features.

3 Proposed methodology

In this study, the methodology for detecting emotions from MIDI musical files begins with extracting features from the dataset, which is crucial for the model's analysis, as shown in Figure 1. These features are fed into a new recognition model called an augmented residual gated recurrent unit. This model is made to accurately detect emotions. A key part of this process is optimizing the GRU's hyper-parameters using the adaptive Red Fox algorithm, enhancing the model's efficiency. The methodology culminates in classifying emotions into four quadrants: positive-high, positive-low, negative-high, and negative-low, allowing for a detailed understanding of the emotional spectrum in the music. This approach ensures precision in interpreting the emotional content of MIDI files, significantly contributing to emotion recognition in music. Enhanced RGRU MIDI musical file Emotion class PH: Positive-high PL: Positive-low NH: Negative-high NL: Negative-low Negative-low Feature extraction Hyperparameters Adaptive Red Fox algorithm.

3.1 Symbolic musical representation

Symbolic musical representation, similar to language modeling, involves converting MIDI files into discrete sequences of notes, mirroring musical events in a format akin to vocabulary (YGhatas et al., 2022). Tools like PrettyMIDI are used to extract specific details, such as each note's pitch, velocity, and duration. These details are then shown visually in Figure 2 using a set of pitch, duration, and hold elements. This is done through one-hot encoding, which turns complicated musical data into a format that is easy to understand. This method not only captures basic note elements but also encompasses key musical structures like melody, harmony, rhythm, and timbre, which are essential for understanding the emotional impact of music (Coutinho and Cangelosi, 2011).

3.1.1 MIDI standard

The Musical Instrument Digital Interface (MIDI) is a symbolic music format that stands apart by recording musical performances using high-level music features, diverging from traditional audio formats that rely on low-level sound features. In MIDI, the focus is on abstractions like musical keys and chord progressions. A MIDI file typically consists of multiple tracks, each capable of independently playing a different instrument, providing a rich, layered musical experience (Sethares et al., 2005). Central to the MIDI format is the concept of the "tick," which serves as the fundamental time unit. This unit is crucial in regulating all aspects of timing in a MIDI file, from the phases of notes to the intervals between them, ensuring a precise and accurate representation of musical timing.

3.2 Music feature extraction

Initially, the MIDI dataset's characteristics are extracted. There are rhythmic characteristics, dynamic characteristics, and statistical aspects.

3.2.1 Harmony features

Musical tones may be utilized to observe harmonics. The spectrogram of monophonic music reveals the harmonics with great clarity (Pickens and Crawford, 2002). In polyphony, where so many instruments and vocalists are used simultaneously, it is challenging to detect harmonics. A method for calculating harmonic distributions is the solution to this conundrum.

$$Hs(f) = \sum_{k=1}^{M_h} \min\left(\left\| S(f) \right\|, \ \left\| S(kf) \right\|\right) \tag{1}$$

Here, $M_h \rightarrow$ the maximum number of harmonics to be considered. The most prevalent incidence of the phenomenon

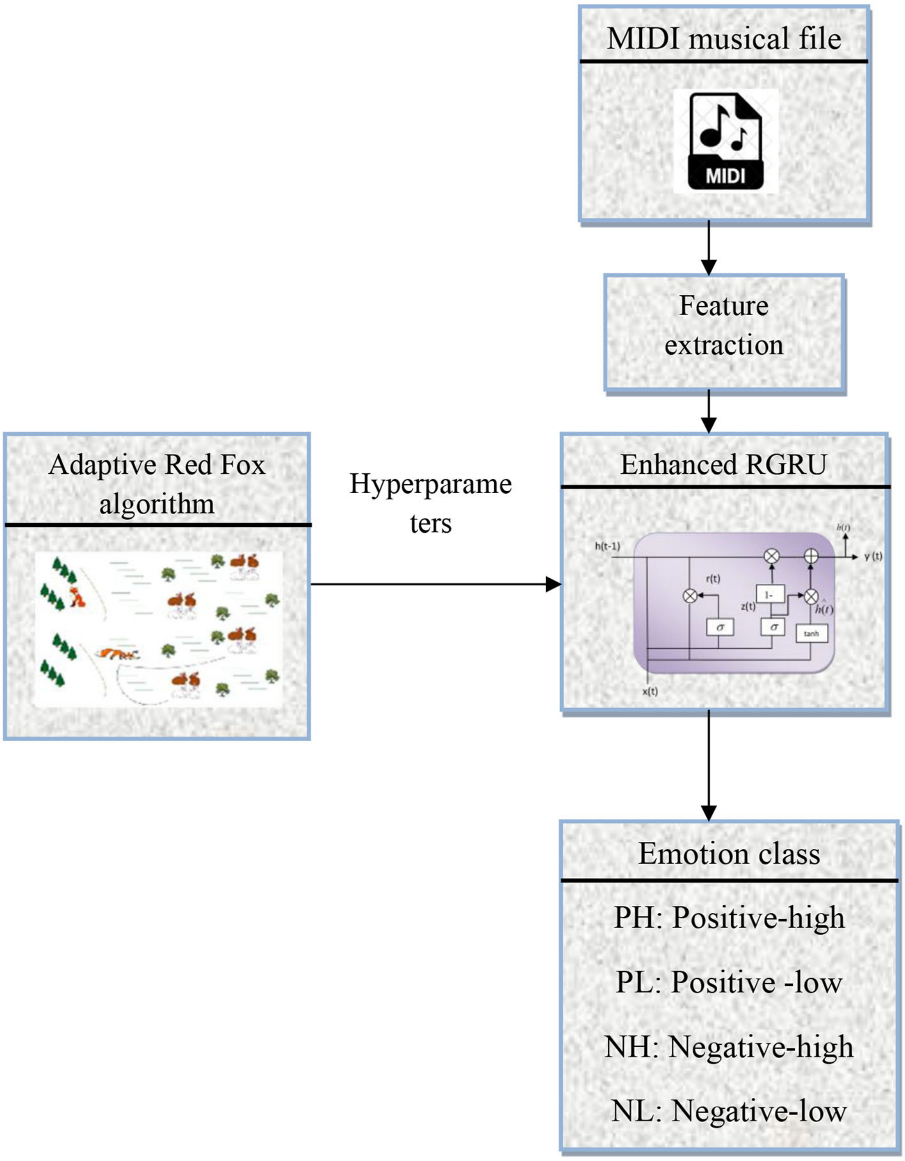


FIGURE 1
Emotion classification workflow for MIDI files.

f → Key frequency

S → The source signal's short-time Fourier transform (STFT).

The min function is applied to the equation so that only the powerful fundamentals and harmonics produce a significant HS value. After calculating the average of each frequency using (1), the standard deviation of each frequency was calculated.

3.2.2 Rhythmic features

Rhythm is a fundamental aspect of music, encompassing key rhythmic characteristics like tempo and cadence, which are essential in defining a musical piece's character. At the heart of musical cadence lies the beat, serving as the primary rhythm indicator. Tempo, a critical component of rhythm, is conventionally measured in beats per minute (BPM). This metric sets the overall rhythmic framework, dictating the speed and flow of the music (Fernández-Sotos et al., 2016). In practice, several techniques are employed to gauge the rate and consistency of these rhythmic pulses. To accurately determine the regularity and pace of the tempo, two main metrics are used: the overall tempo, which is quantified in pulses per minute, and the standard deviation of the intervals between beats. These measures together provide a comprehensive understanding of the tempo's stability and variation, thereby offering insights into the rhythmic structure that underpins the musical composition.

3.2.3 Dynamic features

Dynamics in music are deciphered by examining the pitch salience of every note in relation to others in the composition. Each

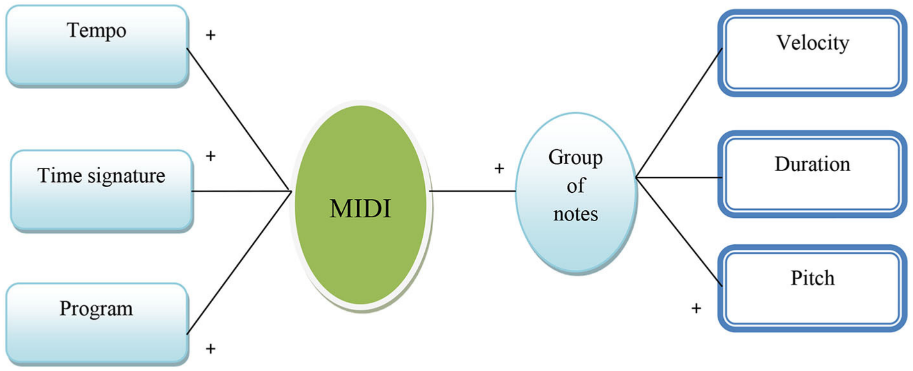


FIGURE 2
Components of MIDI data structure.

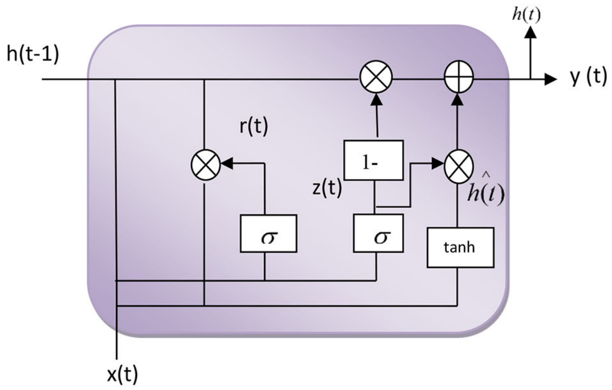


FIGURE 3
The structure of enhanced RGRU's neurons.

note's intensity and its variation are determined by comparing it with the mean and standard deviation of all notes. Consequently, note intensities are classified as high (vigorous), medium, or low (smooth). Dynamic attributes, including RMS Energy, Low Energy Rate, Instantaneous Level, Frequency and Phase, Loudness, Timbral Width, Volume, Sound Balance, Note Intensity Statistics, and Note Intensity Distribution, encapsulate the essence of dynamic levels like forte and piano. Further nuances in dynamics are captured by metrics such as Transition Ratios, Crescendo, and Decrescendo (Panda et al., 2020).

3.3 Enhanced Residual Gated Recurrent Unit architecture

The advanced RGRU, a refined version of the GRU, is depicted in Figure 3, providing a detailed visual representation of its neural architecture. In this study, the recognition model for emotion detection in MIDI files employs an RGRU, into which extracted features are fed. The RGRU is designed to overcome the limitations of traditional GRU models, such as slow convergence and inadequate learning efficacy, particularly in handling complex time series data. Its innovative structure uses feedback from the reset gate to modify the update gate, enhancing the functionality and reducing redundant state information. This modification not only speeds up convergence but also significantly improves the model's learning capacity.

Assuming that the input sequence is $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \boldsymbol{x}_t)$, followed by an update of the gate at t and a reset of the gate, the formula for calculating the standard, enhanced RGRU unit output is as follows:

$$r_t = \sigma\left(V_r * \left[h_{t-1}, x_t\right]\right) \tag{2}$$

$$z_t = \sigma\left(V_z * \left[h_{t-1}, x_t * r_t\right]\right) \tag{3}$$

$$n_t = \tanh\left(V * \left[r_t * h_{t-1}, x_t\right]\right) \tag{4}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * n_t \tag{5}$$

$$y_t = \sigma\left(V_0 * h_t\right) \tag{6}$$

The sigma "σ" typically represents the standard deviation, a measure of the amount of variation or dispersion in a set of values. While "V" represents the MIDI volume or velocity "V0" denotes the initial value of a variable represented by "V".

The formula's symbols $z_t r_t$ have the same significance as standard GRU—the neurons. According to Figure 3, the enhanced RGRU neuron differs from the GRU neuron in that it is multiplied by the previous time at the update gate to conceal the state weight, allowing the reset gate to rescreen the current input data. In other words, the output of the reset gate is used to modify the update gate to optimize the neuron structure and Equation (3), the enhanced RGRU. The neuron structure of the neural network is more logical than that of the GRU; the concealed state at each instant can be made more transparent, and gradient attenuation is moderately reduced. As a consequence, the RGRU was upgraded. The model's learning efficacy and prediction accuracy have improved, and it can maintain a greater dependence on distance information. The deep-enhanced RGRU neural network comprises input, output, and hidden layers. Neurons make up the concealed layers of the RGRU. Refining the GRU model's learning mechanism enhances the recursive transmission of information between neurons and the capacity to retain information.

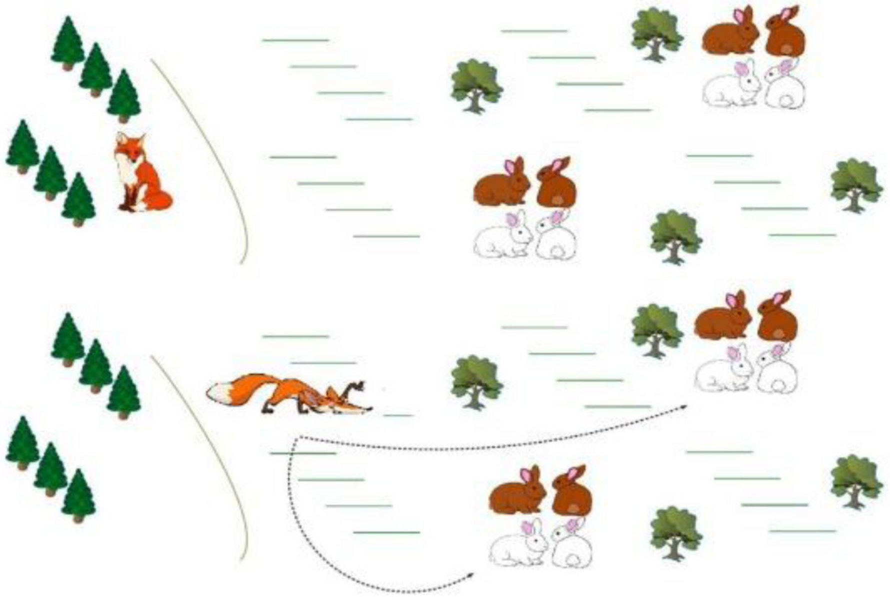


FIGURE 4
Fox and Rabbits interaction simulation.

3.3.1 Adaptive Red Fox algorithm

The ARFA is integrated to fine-tune the RGRU's hyperparameters, drawing inspiration from the hunting behavior of red foxes. This behavior, characterized by searching for prey in snow, is the foundation of the FOX algorithm (Cervený et al., 2011; Mohammed and Rashid, 2023). However, the red fox algorithm tends to converge prematurely, often getting stuck in local optima. To counter this, the Levy Flight method is incorporated, introducing diversity among search agents. This addition helps in avoiding local minima, thus enhancing the overall search efficiency and effectiveness of the algorithm. Consequently, combining the Levy Flight mechanism with the Red Fox algorithm enhances the optimization effectiveness. Figure 4 illustrates the hunting behavior of the red fox.

The procedural steps are as follows:

- In the snow on the ground obstructing the prey's vision, the red fox resorts to random hunting.
- The red fox relies on ultrasonography emitted by the prey to locate it, followed by a period of approach.
- By listening to the prey's sounds and analyzing time intervals, the red fox determines the distance between itself and the prey.
- The establishing the prey's distance, the red fox calculates the required jump, proceeding with random walking based on the shortest distance and optimal position.

The steps involved in Adaptive Red Fox Algorithm are explained as follows:

(i) Initialization

The population, known as the Y matrix, is initially initialized by FOX. Red foxes' positions are represented by a Y. Here, hyperparameters such as GRU. Neurons (G_n) and the learning rate (l) should be considered solutions in this study.

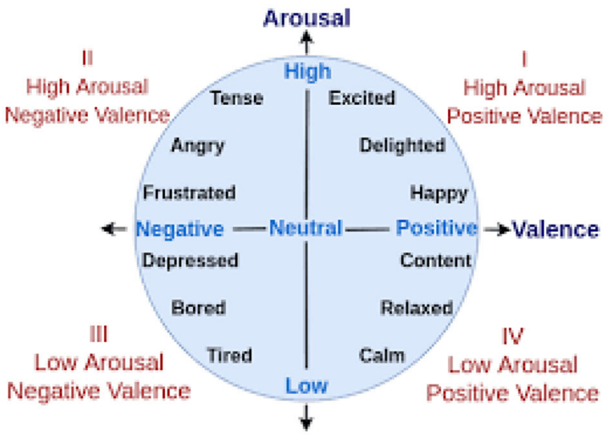


FIGURE 5
Distribution of different emotions on arousal-valence space.

(ii) Fitness function

Then, standard benchmark functions are used to assess the fitness of each search agent after each cycle. In order to find the best fitness () and matching optimal location, we compare the fitness values of individual search agents, represented by rows in an X matrix, to the fitness values of all other agents. The fitness of the previous rowFit_i through the course of iterations is used. Fitness function (Fit_n) can be calculated by using:

Accuracy, *TP* - true positive, *FP*- false positive, *TN*- true negative and *FN*- false negative. These values are crucial for calculating various performance metrics.

$$Fit_n = Max\ \left(Accuracy\right) \tag{7}$$

As an estimation of the accuracy, $Accuracy = \dfrac{TN + TP}{TP + FP + TN + FN}$ (8)

The formula (8) is used to measure the overall correctness of the model. Similarly, the following metrics are also successfully calculated:

Precision: TP/(TP + FP). This indicates the proportion of positive identifications that were actually correct.

Recall (Sensitivity): TP/(TP + FN). This shows the proportion of actual positives that were correctly identified.

***F*-Measure (*F*1 Score):** 2 * (Precision * Recall)/(Precision + Recall). This is the harmonic mean of precision and recall, providing a balance between the two.

(iii) Update the solution

Exploitation Phase: This exploitation stage's random variable value is [0, 1]. Thus, the red fox's updated location must be determined while the random number is more significant than 0.18. Calculate the fox's distance from its prey (dp_{iter}), sound's travel distance (ds_{iter}), and jumping value (j_{iter}) to update its location. To compute the distance between the sound and the red fox (ds_{iter}), use the formula:

$$ds_{iter} = S_{env} * tst_{iter} \tag{9}$$

Sound travels S_{env} at 343 meters per second in the atmospheretst_{iter}, a random value between 0 and 1. The iteration ($iter$) parameter ranges from 1 to 500. The distance between the fox (ds_{iter}) and prey is calculated by halving (dp_{iter}) and is given by:

$$dp_{iter} = ds_{iter} * 0.5 \tag{10}$$

The red fox must move after estimating the distance between it and the prey to jump and seize it. The fox must calculate its jump height (j_{iter}) by:

$$j_{iter} = 0.5 * 9.81 * T^2 \tag{11}$$

To equal the average sound travel time, 9.81 equals gravitational acceleration squared by the jump's up-and-down steps. If a random value between 0 and 1 is more critical than 0.18, the red fox's new location is found using Equations (14) and (15). Only one is executed per iteration due to the p condition. The revised position is calculated using equation (14) if it is more significant than 0.18. Equations determine the current position if the result is less than 0.18 (15). The variable ranges from [0, 0.18] to [0.19, 1]. These values are based on a red fox's leaps toward or away from the northeast. The red fox's new location is estimated using the equation below.

$$Y_{(it+1)} = dp_{iter} * j_{iter} * C_1 \tag{12}$$

$$Y_{(it+1)} = dp_{iter} * j_{iter} * C_2 \tag{13}$$

The value 0.18 was empirically determined to optimize the algorithm's performance, ensuring a balanced and effective search mechanism in the exploitation phase of the Adaptive Red Fox Algorithm. This threshold value is a key factor in the algorithm's ability to accurately and efficiently mimic the strategic hunting pattern of a red fox.

Exploration Phase: During this phase, a fox randomly pursues the best location so far to regulate its random walking. At this stage, the fox could not jump because it had to wander throughout the search area in pursuit of prey. The search is controlled to ensure that the fox wanders randomly to the ideal location using the minimal time variable min T_v and the variable z. Following that, the average time t is determined by dividing T_t by 2. Equations (15) and (16) calculate the min $T_v z$ and variables. Equation (14) can be used to calculate the time transitionT_t.

$$T_t = \frac{sum\left(tst_{iter}\left(i, :\right)\right)}{Dimension} \tag{14}$$

$$\min T_v = Min\left(T_t\right) \tag{15}$$

$$z = 2 * \left(iter - \left(\frac{1}{MaxT_{iter}}\right)\right) \tag{16}$$

Use this method to make sure that the fox checks out the food in a random way. The best answer (Y_{iter}) found significantly affects the exploration phase. The fox's approach to exploring the search space $Y_{(it+1)}$as it looks for a new place to go is shown in Equation (17).

$$Y_{(it+1)} = Y_{it} * rand(1, Dimensiom) * MinT * z \tag{17}$$

Levy flight: When Levy Flight (LF) is implemented, it optimizes the diversity of search agents, ensuring that the algorithm will effectively explore a position while achieving the lowest local avoidance possible.

$$\vec{Y}_{(it+1)} = \vec{Y}_{iter} + \mu\, sign\left[rand - 1/2\right] \oplus levy \tag{18}$$

Here, it represents mean entry-wise multiplication, $\vec{Y}_{iter}$ is the ith Fox location at iteration, μ is a uniformly distributed random value, and finally denotes a random number falling between [0, 1].$sign\left[rand - 1/2\right]$ I only had three values, which were 1, 0, and 1. The Levy Flight produced the following random walk distributions.

$$levy \sim u = t^{-\lambda},\ \ 1 < \lambda \leq 3 \tag{19}$$

Levy flight step lengths s_l are as follows:

$$s_l = \frac{\mu}{|v|^{1/\beta}} \tag{20}$$

λ is constructed using the formula for $\lambda = 1 + \beta$ where$\beta = 1.5\, and\, \mu = N(0, \sigma_{\mu}^{2})$ the identical normal stochastic distributions with

$$\sigma_{\mu} = \left[\frac{\Gamma\,(1+\beta)\, x\, \sin(\pi\, x\, \beta/2)}{\Gamma\,\left(((1+\beta/2))\; x\, \beta\; x 2^{(\beta-1)/2}\right)} \right] \text{and } \sigma_{v} = 1 \tag{21}$$

Incorporating the Levy Flight mechanism into the search process introduces a diversity that allows for a more comprehensive exploration of the solution space, thereby improving the effectiveness of the overall optimization process.

(iv) Termination

The above phases are continued until the optimal solution or optimal weights of RGRU are reached. Otherwise, the algorithm will be terminated. Levy Flight will significantly improve the Red Fox algorithm's search capabilities and protect against local minima.

4 Result and discussion

The implementation of our proposed emotion recognition technique was carried out using Python. In this study, we focused on checking how computationally efficient different processing steps are in the Python environment, such as training and classification. For the assessment, we utilized the Classical Music MIDI dataset, featuring works from nineteen renowned composers, sourced from Piano MIDI. This offered a wide variety of classical piano MIDI files, some of which had audio versions to accompany the playing of the scores. In our methodology, 20% of the dataset was dedicated to evaluating the generation model, with the remaining 80% used for training. Section 4.2 details the performance analysis of the model.

4.1 Dataset analysis

The EMOPIA dataset that was used in this study gives a lot of information about each sample, such as related data, segmentation annotations, and Jensen-Shannon divergence for different emotion quadrant pairs. To facilitate the use of MusPy, MIDI data has been incorporated into the library. However, due to copyright constraints, audio files are not directly released; instead, YouTube links are provided for access. The availability of these songs is subject to the copyright laws of the respective countries and the decisions of the rights holders regarding their availability on the platform.

The study delves into an array of musical elements that are instrumental in shaping the emotions experienced by listeners. The study aims to find out how the different MIDI features are distributed across the four emotional quadrants in order to figure out how these musical features are related to emotions in EMOPIA. The study picks out and shows the most distinguishing features of the different aspects that were looked at, giving us information about the most important parts of music that affect how we feel.

The frequency and intensity of note occurrences serve to measure music arousal, as depicted in Figure 5. This is gauged using three metrics: note length, note density, and note velocity. Note density is the number of notes per beat, and note length is the average duration of a note within a beat. Note velocity, obtained from MIDI data, reflects the strength of each note. These metrics are essential in understanding the music's rhythmic and dynamic properties, corresponding to the emotional states.

4.2 Performance analysis

The developed model in this study aims to categorize emotions into four distinct quadrants: positive-high, positive-low, negative-high, and negative-low. To evaluate its effectiveness, the model employs the EMOPIA dataset. Key performance indicators used for assessment include accuracy, precision, recall, and *F*-score. A big part of this study is comparing how well the new RGRU model works with other models like GRU, LSTM, and CNN. The next part will go into more detail about this comparison by looking at how well the proposed approach works compared to these well-known classification models using a number of different performance metrics.

4.2.1 Performance analysis of positive high quadrant

Figure 6 presents the confusion matrices for the positive high quadrant, summarizing the predictive accuracy of different classification models. The proposed RGRU model correctly identified 157 instances as Positive High and another 120 instances as not belonging to this class. This shows how accurate the model is, with only two false positives and one false negative. This indicates strong model performance with high true positive and true negative rates, coupled with very few misclassifications.

In contrast, the existing GRU model demonstrates slightly diminished accuracy, with 143 true positives and 99 true negatives. It also recorded a higher number of false classifications, with five false positives and three false negatives. The LSTM model follows a similar trend but with more pronounced inaccuracies, tallying 132 true positives and 110 true negatives, alongside 10 false positives and eight false negatives. With only 117 true positives and 110 true negatives, the DNN model is the most different from the proposed RGRU model. It also has the highest error rate, with 13 false positives and 10 false negatives. Overall, the RGRU model does better than the others because it consistently makes more correct predictions and fewer mistakes. This suggests that it is the most reliable model for identifying the Positive High quadrant in this study.

Figure 7 shows the ROC and AUC graphs for the positive high quadrant. The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1-FPR). Classifiers that give curves closer to the top-left corner indicate better performance. The AUC provides an aggregate measure of performance across all possible classification thresholds. It is the area under the ROC curve, with a value between 0 and 1. A model with perfect predictive accuracy would have an AUC of 1, meaning it has a good measure of

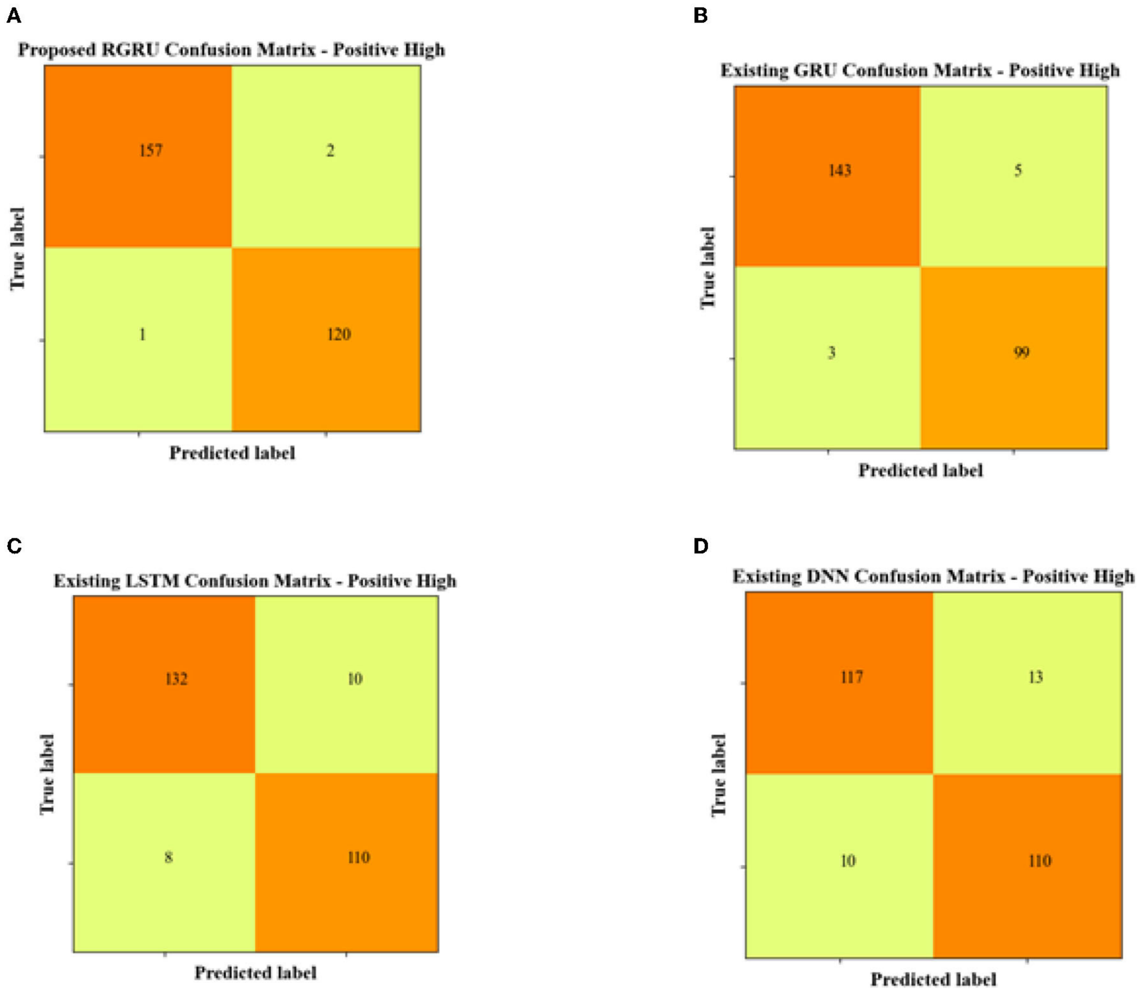


FIGURE 6
Confusion matrix for positive high. **(A)** Proposed RGRU. **(B)** Existing GRU. **(C)** Existing LSTM. **(D)** Existing DNN.

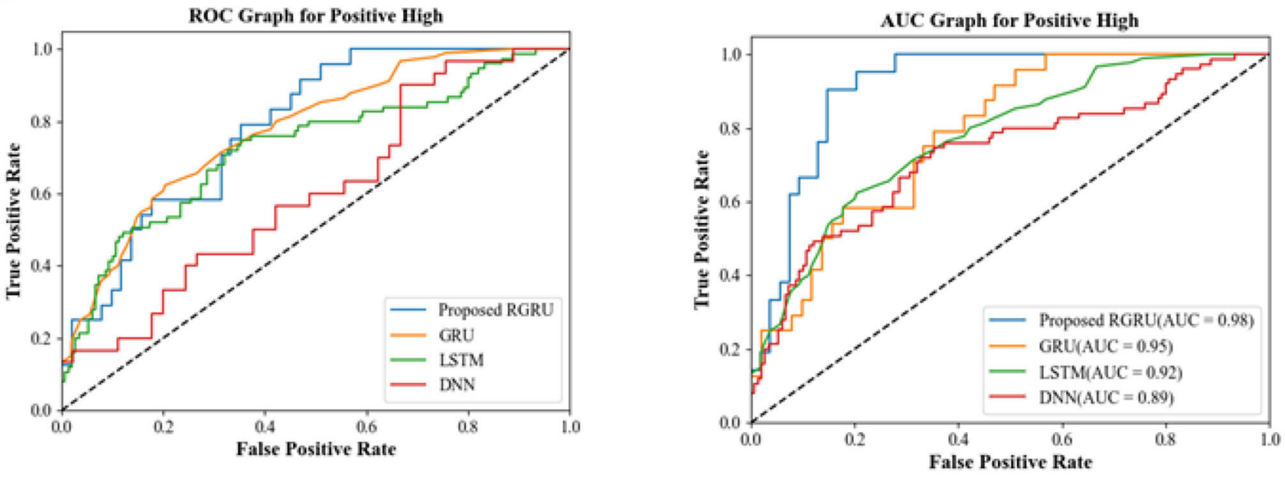


FIGURE 7
R.O.C. and A.U.C. graph for positive high.

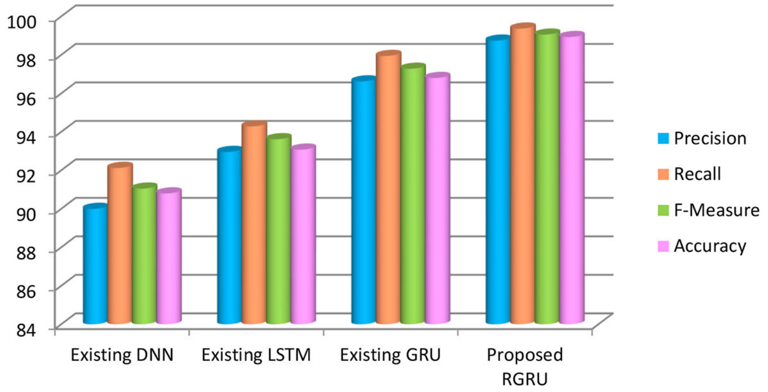


FIGURE 8
Precision, recall, *F*-measure and accuracy-based analysis of various models (positive high).

TABLE 1 The performance evaluation of various models (positive high).

Metric \ Method	Existing DNN	Existing LSTM	Existing GRU	Proposed RGRU
Precision	90	92.95775	96.62162	98.74214
Recall	92.12598	94.28571	97.94521	99.36709
F-measure	91.05058	93.61702	97.27891	99.05363
Accuracy	90.8	93.07692	96.8	98.92857

separability. A model with no discriminative power has an AUC of 0.5, meaning it does as well as random chance.

From the research work, for the positive high quadrant, the proposed RGRU has the highest AUC, indicating it outperforms the other models in distinguishing between the positive high class and the not-positive high class. Existing GRU performs better than LSTM and DNN, but the proposed RGRU outperforms them both. The existing LSTM has a lower AUC than RGRU and GRU but is higher than DNN, suggesting moderate performance. Likewise, the existing DNN has the lowest AUC, indicating the least performance in comparison to the other models. The ROC and AUC graphs demonstrate that the proposed RGRU model has a superior ability to classify the positive high quadrant with more accuracy than the other models.

To see how well the suggested RGRU method works, we look at important performance indicators like accuracy, precision, recall, *F*-measure, and more, which can be seen in Figure 8 and Table 1. We check how well this method works with the Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and Deep Neural Network (DNN) classification methods. With a maximal accuracy of 98.92%, the RGRU technique significantly outperforms alternative models, including LSTM (5.75%), GRU (2.18%), and DNN (8.12%). When compared to the alternative methods, the RGRU strategy exhibits superior performance, as evidenced by its *F*-measure (99.05%), precision (98.74%), and recall (99.36%). The results obtained from the enhanced adaptive red fox algorithm (ARFA) are superior to those obtained from alternative methods, specifically when it comes to identifying positive high-phase emotions, as demonstrated by the table and graph depicted. This serves as a demonstration of how the proposed approach surpasses the present condition of affairs.

4.2.2 Performance analysis of positive low quadrant

The confusion matrix data for the Positive Low quadrant show that the proposed RGRU model is the most accurate, with 130 true positives and 120 true negatives. This shows that it is very good at correctly classifying instances. With only two false positives and a single false negative, it demonstrates remarkable precision in detection. Comparatively, the existing GRU model identified 124 true positives and 119 true negatives but had slightly more misclassifications, with seven false positives and three false negatives. The LSTM model registered 120 true positives and 115 true negatives, with its accuracy further diminished by 10 false positives and 8 false negatives. The DNN model matched the LSTM in true positives but fell behind with only 110 true negatives, and it exhibited the highest error rates, having misclassified 13 false positives and 10 false negatives. Overall, the RGRU model's superior performance is evidenced by its higher correct classifications and minimal errors, affirming its effectiveness in the positive low quadrant compared to the GRU, LSTM, and DNN models.

Figure 9 shows the confusion matrix for the positive low quadrant, and Figure 10 shows the ROC and AUC graph for the positive low quadrant.

The ROC and AUC graphs for the positive low quadrant provide insightful measures of model performance. The ROC graph illustrates the balance between sensitivity and specificity,

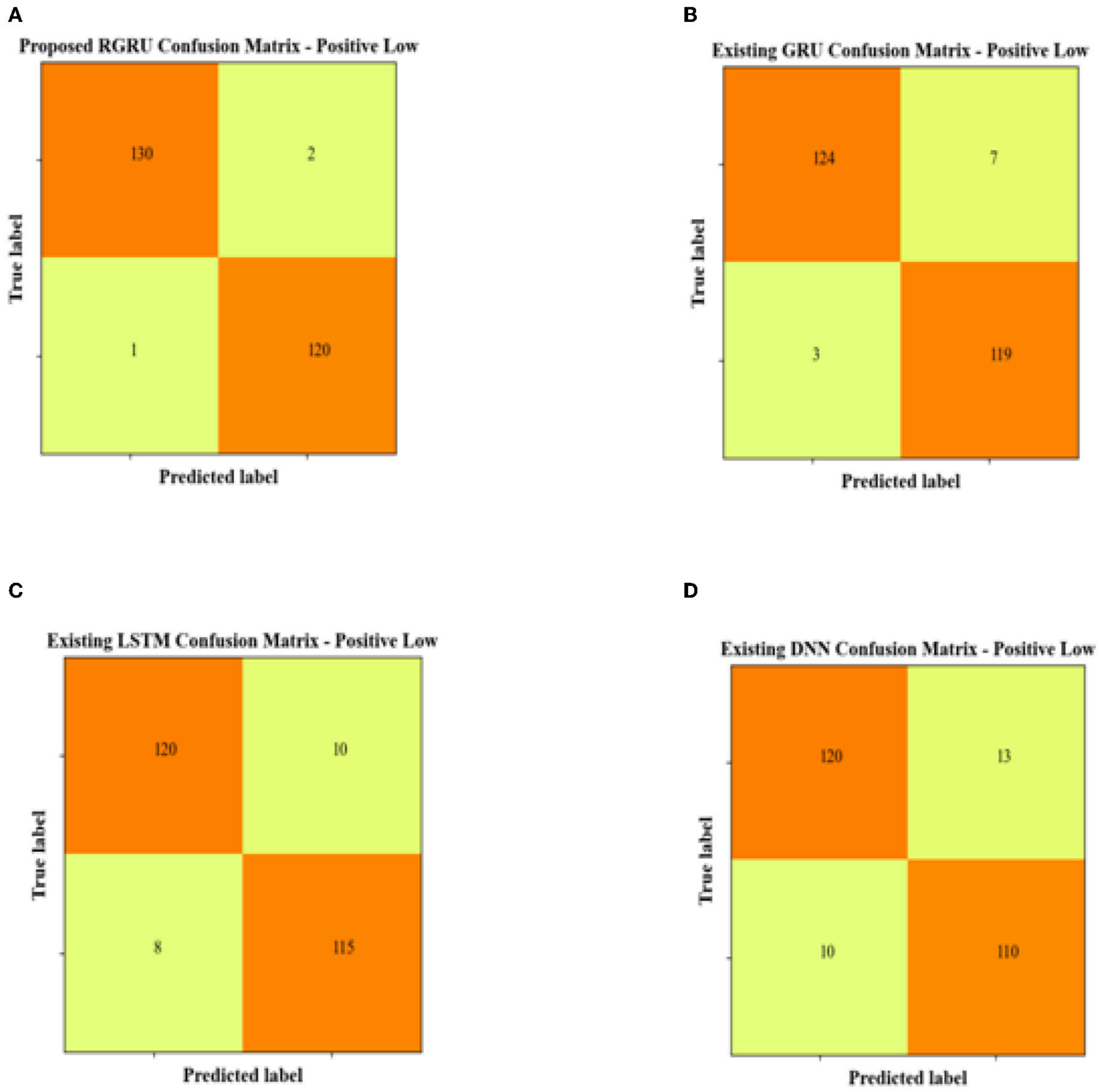


FIGURE 9
Confusion matrix for positive low. **(A)** Proposed RGRU. **(B)** Existing GRU. **(C)** Existing LSTM. **(D)** Existing DNN.

with the proposed RGRU model's curve approaching the ideal top-left corner more closely than the others, signaling its superior performance in correctly identifying true positives while minimizing false positives. The curves of the existing GRU, LSTM, and DNN models don't show this optimal balance as clearly. They lie below that of the RGRU, which means they don't make the trade-off as well.

With an AUC of 0.98, the proposed RGRU model gets the highest score in the AUC graph, which shows how well it does across all possible classification thresholds. This shows that it is very good at telling the difference between classes. The existing models, on the other hand, have lower AUC values-—0.95 for GRU, 0.92 for LSTM, and 0.89 for DNN—which means they are less accurate at classifying things. Together, these AUC scores support what the confusion matrices and the ROC graph showed: the proposed RGRU model is better than the current GRU, LSTM, and DNN models at classifying the Positive Low quadrant with more accuracy and a better ability to tell the classes apart.

In Figure 11 and Table 2, the study assesses the effectiveness of the proposed RGRU method using metrics such as precision, recall, *F*-measure, and accuracy. This method is compared against three alternative classification methods: GRU, LSTM, and DNN. The results shown in Figure 11 show that the RGRU method is more accurate than GRU by 2.77%, LSTM by 5.93%, and DNN by 7.91%. This demonstrates that the proposed RGRU method achieves the highest accuracy among the compared methods. Furthermore, the proposed approach also records the highest precision at 98.48%, recall at 99.23%, and *F*-measure at 98.85%. The table unmistakably demonstrates that the RGRU, with the Adaptive Red Fox Algorithm (ARFA) enhancement, performs better than the other techniques, especially in the positive low phase of emotion recognition. This

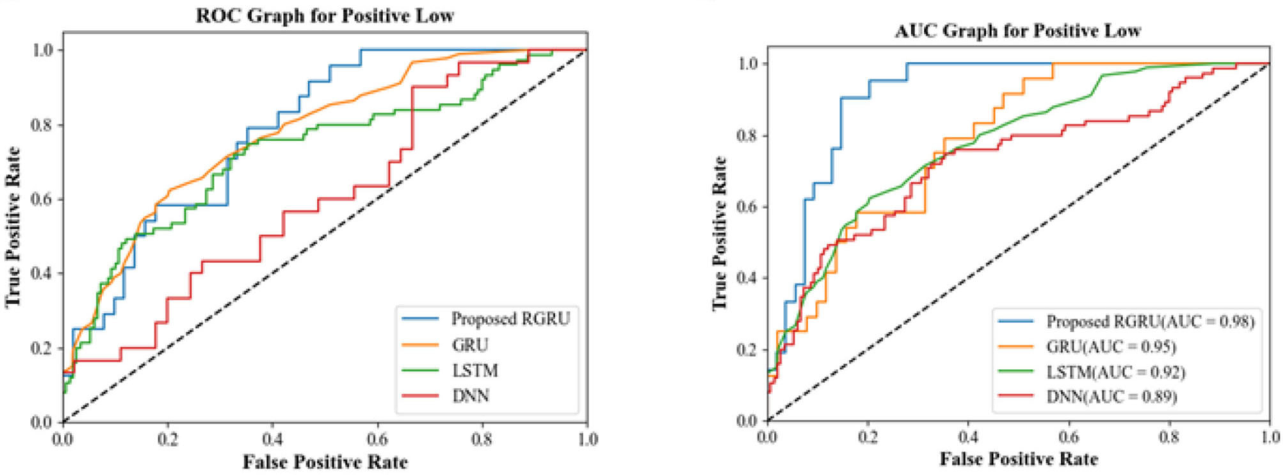


FIGURE 10
R.O.C. and A.U.C. graph for positive low.

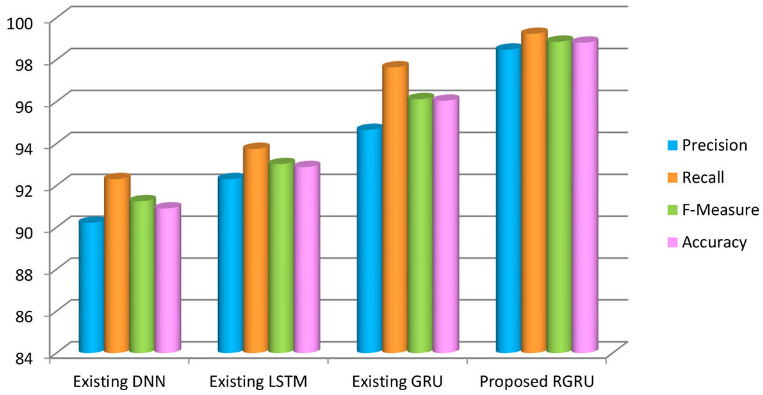


FIGURE 11
Precision, recall, *F*-measure and accuracy-based analysis of various models (positive low).

TABLE 2 The performance evaluation of various models (positive low).

Method / Metric	Existing DNN	Existing LSTM	Existing GRU	Proposed RGRU
Precision	90.22556	92.30769	94.65649	98.48485
Recall	92.30769	93.75	97.6378	99.23664
F-measure	91.25475	93.02326	96.12403	98.85932
Accuracy	90.90909	92.88538	96.04743	98.81423

highlights the effectiveness of the RGRU method in outperforming competing approaches.

4.2.3 Performance analysis of negative high quadrant

The proposed RGRU model stands out in the Negative High confusion matrix with 154 true positives, which correctly identify Negative High instances, and 120 true negatives, which correctly identify non-Negative High instances. It demonstrates a robust classification capability with only three false positives and an equal number of false negatives.

The current GRU model has a higher count of 187 true positives, but it is less accurate with eight false positives and five false negatives, which means that wrong classifications are more likely to happen. On the other hand, the LSTM model, with 162 true positives and 120 true negatives, also displays an increased rate of misclassification, as evidenced by 18 false positives and

10 false negatives, which underscores a potential compromise in model reliability.

The DNN model, despite having a commendable number of 130 true negatives, falls short in accuracy, with the lowest true positive count at 145 and the highest false positive count at 25, accompanied by 10 false negatives. This indicates a substantial reduction in its efficacy in the negative high quadrant compared to the RGRU model.

In essence, the RGRU model's performance in the negative high quadrant surpasses that of the GRU, LSTM, and DNN models, as reflected by its higher correct predictions and lower misclassifications, showcasing its effectiveness and reliability in emotion classification within this specific context.

Figure 12 shows the confusion matrix for the negative high quadrant, and Figure 13 shows the ROC and AUC graph for the negative high quadrant. The proposed RGRU model demonstrates superior proficiency, with its curve nearing the top-left corner, an indication of an excellent balance between sensitivity and specificity. In comparison, the curves representing the GRU, LSTM, and DNN models are positioned lower, signifying a less optimal trade-off and reduced effectiveness in distinguishing the negative high class.

AUC, measures a model's accuracy over a broad range of threshold values. The RGRU model's AUC value of 0.98 indicates that it has a significant ability to differentiate classes. The classification performance of the DNN model is 0.89, whilst the LSTM model shows a performance of 0.92. With an AUC of 0.95, however, the GRU model performs better than both.

To assess the effectiveness of the strategy depicted in Figure 14 and Table 3, accuracy, precision, recall, and the *F*-measure are used. Using these measures, we compare our proposed RGRU technique against three distinct classification strategies: GRU, LSTM, and DNN. Our technique has a maximum accuracy of 98.06%, outperforming GRU by 2.26%, LSTM by 7.1%, and DNN by 9.36%. Our technique's improved performance is also visible in other parameters, such as recall (98.24%), precision (96.39%), and *F*-measure (98.39%). Table 3 shows the results of the public inspection technique that we proposed. The results show that our strategy was effective throughout the negative high phase; the improved performance of the RGRU is attributed to the Adaptive Red Fox Algorithm (ARFA). This contrast highlights the enormous advances that our proposed methodology offers to the problem of emotion categorization.

4.2.4 Performance analysis of negative low quadrant

The proposed RGRU model's confusion matrix does a great job in the negative low quadrant, with a high number of true positives (169) and true negatives (90). This means that the classification is correct, with only a few cases being wrongly labeled (false positives at 3) or missed (false negatives also at 3). This suggests a precise model for identifying negative and low emotions. The current GRU model, on the other hand, has a slightly lower level of accuracy than the RGRU model, with 162 true positives and 90 true negatives. This is because it has more false positives (8) and false negatives (5).

Increased misclassifications, with 136 true positives and 100 true negatives, but also a noticeable rise in both false positives (19) and false negatives (10), highlight the LSTM model's further decreased performance and suggest that it is less reliable for accurate classification of negative low emotions. The DNN model presents the lowest performance in the group, with the lowest count of true positives (121) and the highest count of false negatives (15), coupled with a considerable number of false positives (19). The DNN model's confusion matrix clearly illustrates its challenges in accurately classifying negative emotions, with considerable room for improvement in its predictive capabilities.

Figure 15 shows the confusion matrix for the negative low quadrant, and Figure 16 shows the ROC and AUC graph for the negative low quadrant. For the negative low quadrant, the ROC and AUC graphs provide insightful measures of each model's performance.

The ROC graph for the proposed RGRU model showcases an optimal balance between the true negative rate and the false negative rate, with its curve being the closest to the ideal top-left corner. This shows a better ability to tell the difference between Negative Low and other classes without labeling instances that aren't Negative Low as Negative Low by accident. The ROC curves for the GRU, LSTM, and DNN models are farther from the ideal point, which means that their balance between sensitivity and specificity is not as good. These models have a lower true negative rate for any given false negative rate, signifying a reduced ability to accurately classify negative emotions.

The AUC value for the RGRU model stands at 0.98, the highest among the models, demonstrating its outstanding overall classification performance. This high AUC value means that the RGRU model has a good chance of correctly identifying any given case as either negative or not, for all thresholds. The GRU, LSTM, and DNN models have lower AUC scores (0.95, 0.92, and 0.89, respectively), which means they can't tell the difference between negative low and non-negative low classes as well-across all thresholds. The RGRU model, which performs better at classifying negative low emotions than the GRU, LSTM, and DNN models, supports the confusion matrix results according to the ROC and AUC graphs. The RGRU model is better because it is closer to the ideal points on the graphs, has higher true negative rates, and has a higher AUC value. These factors show that it is better at telling the difference between negative and low emotions and overall performance.

Figure 17 and Table 4 show how well the suggested RGRU-based method works by checking its precision, recall, *F*-measure, and accuracy. This evaluation involves a comparative analysis with other classification algorithms, namely GRU, LSTM, and DNN. Our suggested method performs much better than the others, with a maximum accuracy of 97.73%, which is 2.64% higher than GRU, 8.68% higher than LSTM, and 10.57% higher than DNN. This makes it the most accurate method we looked at.

Furthermore, the suggested method performs exceptionally well in important measures, with an *F*-measure of 98.35%, a precision rate of 97.25%, and a recall rate of 98.25%. As detailed in Table 4, these statistics prominently showcase the method's superior performance compared to other techniques. Particularly noteworthy is the method's performance in the negative low

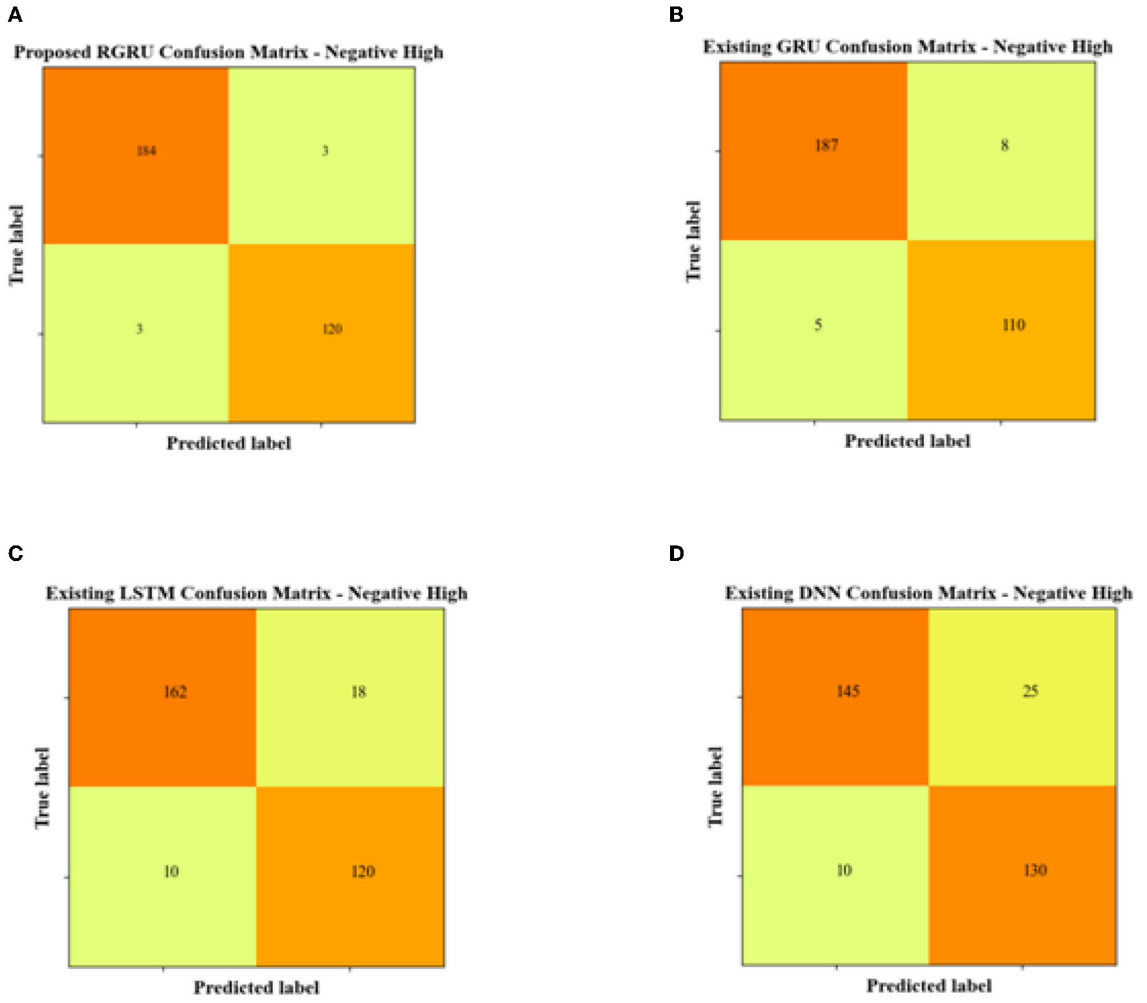


FIGURE 12
Confusion matrix for negative high. **(A)** Proposed RGRU. **(B)** Existing GRU. **(C)** Existing LSTM. **(D)** Existing DNN.

phase, as highlighted in the results section. We think this better performance is because the Adaptive Red Fox Algorithm (ARFA) was used to improve the RGRU. This makes it much better at classifying emotions. I made the right choice by using both RGRU and ARFA in my research. Together, they help reach the goals of the study, specifically by making emotion recognition from MIDI files more accurate and reliable.

The combination of these advanced methods aligns perfectly with the research objectives of accurately identifying and classifying emotions in MIDI musical files. The RGRU's architecture is well-suited for the sequential and temporal nature of music data, while ARFA ensures that the model operates at its highest potential. They work well together to show that these methods are complete and accurate for detecting emotions in MIDI files, proving that they are good for the research goals. The MIDI dataset used in this study appears reliable, as MIDI files accurately encode detailed musical information crucial for emotion recognition. The study's results were checked using statistical methods like *F*-score, accuracy, precision, and recall to measure the model's performance in a quantitative way. We found these results to be even more important by comparing them to results from well-known models like GRU, LSTM, and DNN. This showed that the new model was better at recognizing emotions from MIDI files.

5 Conclusion

This study has introduced a novel approach to discerning the emotional nuances embedded within each MIDI composition, utilizing the enhanced RGRU architecture for hyperparameter optimization through ARFA. We used the EMOPIA dataset and performance metrics like precision, *F*-measure, recall, and accuracy to do a full evaluation of our proposed method to see how well it worked. In the comparative analysis against the existence prediction models, including GRU, LSTM, and DNN, the proposed approach consistently outperformed them in all four quadrants: positive-high (98.92%), positive-low (98.91%), negative-high (98.06%), and negative-low (97.73%). These results underscore our

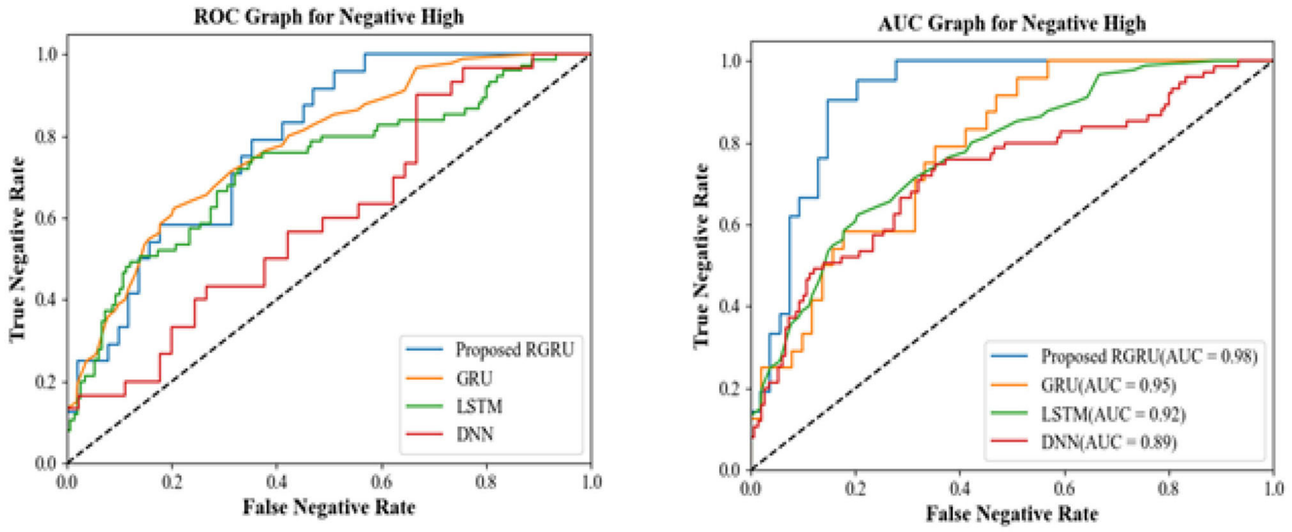


FIGURE 13
R.O.C. and A.U.C. graph for negative high.

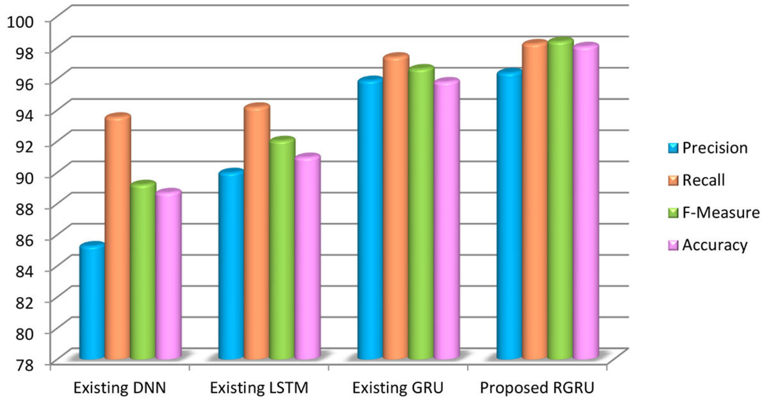


FIGURE 14
Precision, recall, *F*-measure and accuracy-based analysis of various models (negative high).

TABLE 3 The performance evaluation of various models (negative high).

Method / Metric	Existing DNN	Existing LSTM	Existing GRU	Proposed RGRU
Precision	85.29412	90	95.89744	96.39572
Recall	93.54839	94.18605	97.39583	98.24572
F-measure	89.23077	92.04545	96.64083	98.39572
Accuracy	88.70968	90.96774	95.80645	98.06452

innovative methodologies' superior predictive accuracy and overall efficacy.

While emotion recognition in music is a recognized field, its specific application to MIDI compositions is relatively less explored. This research adds originality by focusing on analyzing emotions in MIDI data, which can have unique challenges compared to other audio formats.

The research relies on the EMOPIA dataset for evaluation. If this dataset has biases or limitations regarding diversity and representation of musical emotions, it can impact the generalizability of the findings. The study demonstrates

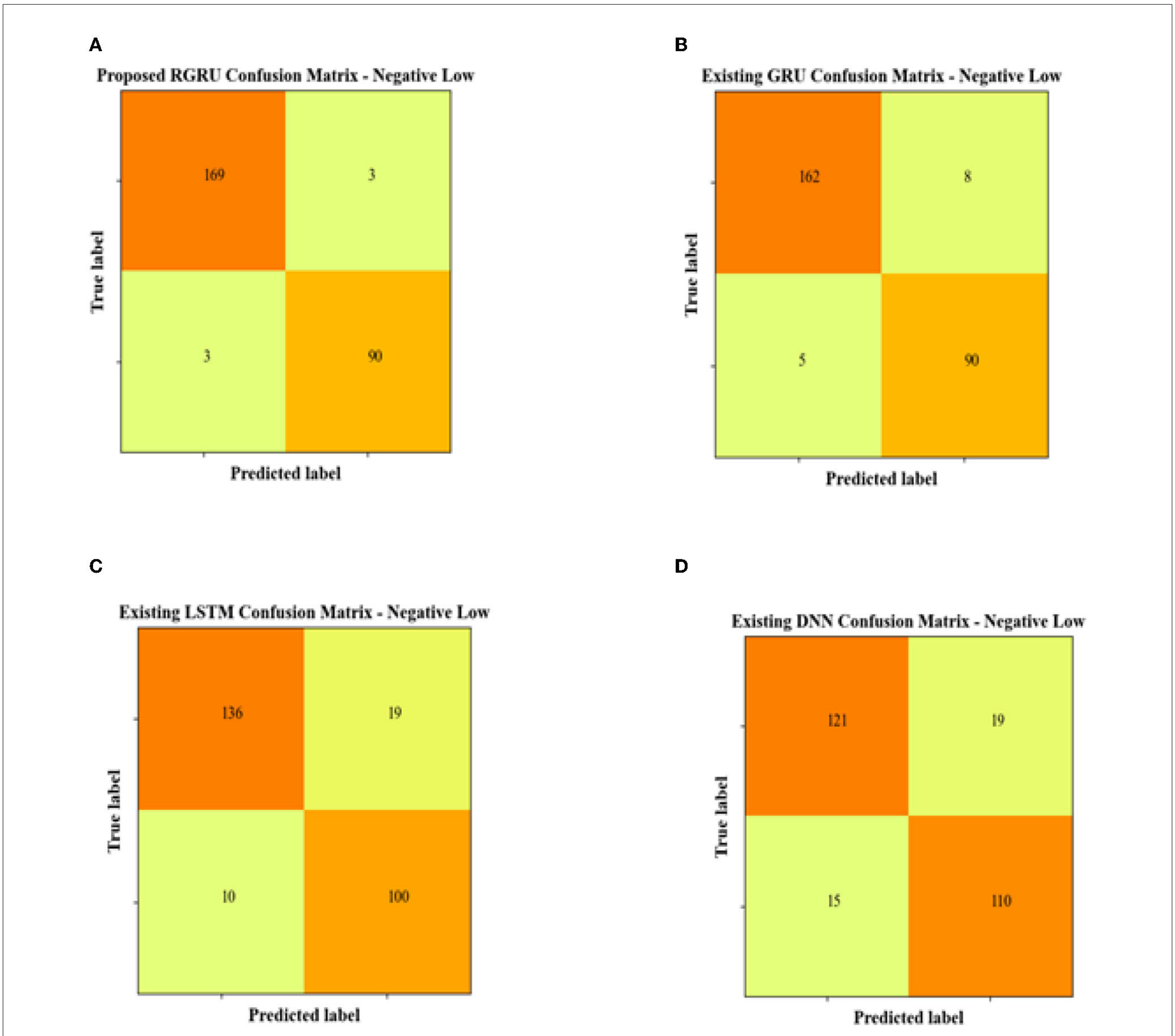


FIGURE 15
Confusion matrix for negative low. **(A)** Proposed RGRU. **(B)** Existing GRU. **(C)** Existing LSTM. **(D)** Existing DNN.

TABLE 4 The performance evaluation of various models (negative low).

Method Metric	Existing DNN	Existing LSTM	Existing GRU	Proposed RGRU
Precision	86.42857	87.74194	95.29412	97.25581
Recall	88.97059	93.15068	97.00599	98.25581
F-measure	87.68116	90.36545	96.14243	98.35581
Accuracy	87.16981	89.0566	95.09434	97.73585

the effectiveness of the proposed approach, but it may not necessarily generalize well to different music genres, styles, or cultural contexts. It's important to acknowledge the scope of its applicability. Emotion recognition in music is inherently subjective. The model's interpretation of emotions might not fully capture the individual listener's experience, potentially leading to discrepancies between the model's classifications and human perception.

Future research should aim to diversify datasets for broader genre coverage, develop algorithms for nuanced emotion detection, ensure hardware scalability, and refine emotion classification methods. These steps will enhance the model's accuracy and

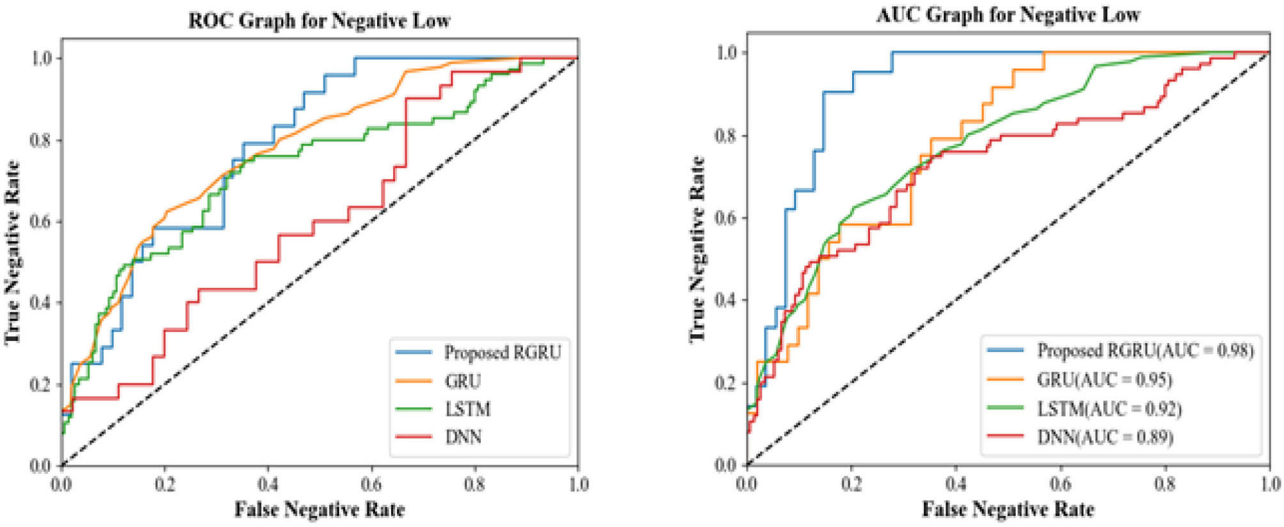


FIGURE 16
R.O.C. and A.U.C. graph for negative low.

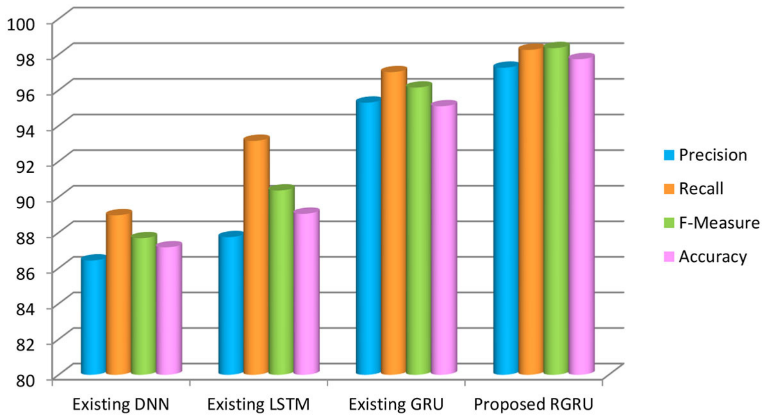


FIGURE 17
Precision, recall, *F*-measure and accuracy-based analysis of various models (negative low).

applicability in diverse musical and cultural settings, ensuring its effectiveness in real-world scenarios.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

VK: Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. MK: Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abboud, R., and Tekli, J. (2020). Integrating nonparametric fuzzy classification with an evolutionary-developmental framework to perform music sentiment-based analysis and composition. *Soft Comp.* 24, 9875–9925. doi: 10.1007/s00500-019-04503-4

Bhatti, A. M., Majid, M., Anwar, S. M., and Khan, B. (2016). Human emotion recognition and analysis in response to audio music using brain signals. *Comput. Human Behav.* 65, 267–275. doi: 10.1016/j.chb.2016.08.029

Bresin, R., and Friberg, A. (2000). The emotional colouring of computer-controlled music performances. *Comp. Music J.* 24, 44–63. doi: 10.1162/014892600559515

Cervený, J., Begall, S., Koubek, P., Nováková, P., and Burda, H. (2011). Directional preference may enhance hunting accuracy in foraging foxes. *Biol. Lett.* 7, 355–357. doi: 10.1098/rsbl.2010.1145

Chen, S. H., Lee, Y. S., Hsieh, W. C., and Wang, J. C. (2015). "Music emotion recognition using deep Gaussian process," in *2015, the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (Taiwan: IEEE), 495–498.

Coutinho, E., and Cangelosi, A. (2011). Musical emotions: predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion* 11, 921. doi: 10.1037/a0024700

Fernández-Sotos, A., Fernández-Caballero, A., and Latorre, J. M. (2016). Influence of tempo and rhythmic unit in musical emotion regulation. *Front. Comput. Neurosci.* 10, 80. doi: 10.3389/fncom.2016.00080

Ferreira, L. N., and Whitehead, J. (2021). Learning to generate music with sentiment. *arXiv* [06125].

Ghatas, Y., Fayek, M., and Hadhoud, M. (2022). A hybrid deep learning approach for musical difficulty estimation of piano symbolic music. *Alexandria Eng. J.* 61, 10183–10196. doi: 10.1016/j.aej.2022.03.060

Good, M. (2001). MusicXML for notation and analysis. *Virt. Score* 12, 113–124.

Hosken, D. (2014). *An Introduction to Music Technology*. London: Routledge.

Hsu, Y. L., Wang, J. S., Chiang, W. C., and Hung, C. H. (2017). Automatic ECG-based emotion recognition in music listening. *IEEE Transact. Affect. Comp.* 11, 85–99. doi: 10.1109/TAFFC.2017.2781732

Hung, H. T., Ching, J., Doh, S., Kim, N., Nam, J., and Yang, Y. H. (2021). EMOPIA: a multi-modal pop piano dataset for emotion recognition and emotion-based music generation. *arXiv*.

Juslin, P. N., and Timmers, R. (2010). "Expression and communication of emotion in music performance," in *Handbook of Music and Emotion: Theory, Research, Applications* (Washington, DC: Oxford University Press), 453–489.

Koh, E., and Dubnov, S. (2021). "Comparison and analysis of deep audio embeddings for music emotion recognition," *AAAI Workshop on Affective Content Analysis*. New York, NY: Cornell University.

Krumhansl, C. L. (2002). Music: a link between cognition and emotion. *Curr. Dir. Psychol. Sci.* 11, 45–50. doi: 10.1111/1467-8721.00165

Li, B., Liu, X., Dinesh, K., Duan, Z., and Sharma, G. (2018). Creating a multitrack classical music performance dataset for multi-modal music analysis: challenges, insights, and applications. *IEEE Transact. Multimedia* 21, 522–535. doi: 10.1109/TMM.2018.2856090

Luck, G., Toiviainen, P., Erkkilä, J., Lartillot, O., Riikkilä, K., Mäkelä, A., et al. (2008). Modelling the relationships between emotional responses to and the musical content of music therapy improvisations. *Psychol. Music* 36, 25–45. doi: 10.1177/0305735607079714

Ma, L., Zhong, W., Ma, X., Ye, L., and Zhang, Q. (2022). Learning to generate emotional music correlated with music structure features. *Cogn. Comp. Syst.* 4, 100–107. doi: 10.1049/ccs2.12037

Malik, M., Adavanne, S., Drossos, K., Virtanen, T., Ticha, D., and Jarina, R. (2017). Stacked convolutional and recurrent neural networks for music emotion recognition. *arXiv*. doi: 10.23919/EUSIPCO.2017.8081505

Modran, H. A., Chamunorwa, T., Ursuţiu, D., Samoilă, C., and Hedeşiu, H. (2023). Using deep learning to recognize therapeutic effects of music based on emotions. *Sensors* 3, 986. doi: 10.3390/s23020986

Mohammed, H., and Rashid, T. (2023). FOX: a FOX-inspired optimization algorithm. *Appl. Intell.* 53, 1030–1050. doi: 10.1007/s10489-022-03533-0

Nanayakkara, S. C., Wyse, L., Ong, S. H., and Taylor, E. A. (2013). Enhancing the musical experience for people who are deaf or hard of hearing using visual and haptic displays. *Hum. Comp. Interact.* 28, 115–160.

Nienhuys, H.-W., and Nieuwenhuizen, J. (2003). "LilyPond, a system for automated music engraving," in *Proceedings of the XIV Colloquium on Musical Informatics (XIV CIM 2003), Vol. 1* (Switzerland: Citeseer), 167–171.

Panda, R., Malheiro, R., and Paiva, R. P. (2018). Novel audio features for music emotion recognition. *IEEE Transact. Affect. Comp.* 11, 614–626. doi: 10.1109/TAFFC.2018.2820691

Panda, R., Malheiro, R. M., and Paiva, R. P. (2020). Audio features for music emotion recognition: a survey. *IEEE Transact. Affect. Comp.*

Pickens, J., and Crawford, T. (2002). "Harmonic models for polyphonic music retrieval," in *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (University of London), 430–437.

Renz, K. (2002). *Algorithms and Data Structures for a Music Notation System Based on Guido Music Notation* (Ph.D. thesis), Darmstadt: Technische Universita.

Sethares, W. A., Morris, R. D., and Sethares, J. C. (2005). Beat tracking of musical performances using low-level audio features. *IEEE Transact. Speech Audio Process.* 13, 275–285. doi: 10.1109/TSA.2004.841053

Shou, L., Mao, K., Luo, X., Chen, K., Chen, G., and Hu, T. (2013). "Competence-based song recommendation," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Zhejiang University, China), 423–432.

YGhatas, S., Fayek, M. B., and Hadhoud, M. M. (2022). Generic symbolic music labeling pipeline. *IEEE Access* 10, 76233–76242. doi: 10.1109/ACCESS.2022.3192462

TYPE Original Research
PUBLISHED 02 February 2024
DOI 10.3389/fpsyg.2024.1353022

Artificial intelligence and social intelligence: preliminary comparison study between AI models and psychologists

Nabil Saleh Sufyan[1], Fahmi H. Fadhel[2]*, Saleh Safeer Alkhathami[1] and Jubran Y. A. Mukhadi[1]

[1]Psychology Department, College of Education, King Khalid University, Abha, Saudi Arabia,
[2]Psychology Program, Social Science Department, College of Arts and Sciences, Qatar University, Doha, Qatar

Check for updates

OPEN ACCESS

EDITED BY
Mariacarla Martí-González,
University of Valladolid, Spain

REVIEWED BY
Isabella Poggi,
Roma Tre University, Italy
Juan Luis Martín Ayala,
Universidad Europea del Atlántico, Spain

*CORRESPONDENCE
Fahmi H. Fadhel
✉ fahmi@qu.edu.qa;
✉ fahmi4n@yahoo.com

RECEIVED 09 December 2023
ACCEPTED 22 January 2024
PUBLISHED 02 February 2024

CITATION
Sufyan NS, Fadhel FH, Alkhathami SS and Mukhadi JYA (2024) Artificial intelligence and social intelligence: preliminary comparison study between AI models and psychologists. *Front. Psychol.* 15:1353022.
doi: 10.3389/fpsyg.2024.1353022

Background: Social intelligence (SI) is of great importance in the success of the counseling and psychotherapy, whether for the psychologist or for the artificial intelligence systems that help the psychologist, as it is the ability to understand the feelings, emotions, and needs of people during the counseling process. Therefore, this study aims to identify the Social Intelligence (SI) of artificial intelligence represented by its large linguistic models, "ChatGPT; Google Bard; and Bing" compared to psychologists.

Methods: A stratified random manner sample of 180 students of counseling psychology from the bachelor's and doctoral stages at King Khalid University was selected, while the large linguistic models included ChatGPT-4, Google Bard, and Bing. They (the psychologists and the AI models) responded to the social intelligence scale.

Results: There were significant differences in SI between psychologists and AI's ChatGPT-4 and Bing. ChatGPT-4 exceeded 100% of all the psychologists, and Bing outperformed 50% of PhD holders and 90% of bachelor's holders. The differences in SI between Google Bard and bachelor students were not significant, whereas the differences with PhDs were significant; Where 90% of PhD holders excel on Google Bird.

Conclusion: We explored the possibility of using human measures on AI entities, especially language models, and the results indicate that the development of AI in understanding emotions and social behavior related to social intelligence is very rapid. AI will help the psychotherapist a great deal in new ways. The psychotherapist needs to be aware of possible areas of further development of AI given their benefits in counseling and psychotherapy. Studies using humanistic and non-humanistic criteria with large linguistic models are needed.

KEYWORDS
artificial intelligence, social intelligence, psychologists, ChatGPT, Google Bard, Bing

1 Introduction

Machines have influenced human evolution. The characteristics of each era have been shaped by the tools developed since the First Industrial Revolution (1760–1840), for example, the use of steam machines instead of manual labor, and the Second Industrial Revolution (1870–1914), represented by the use of energy. The use of electricity instead of steam power led to the Third Industrial Revolution (1950–1970), where electronic and communication devices such as computers and portable devices appeared. Today we are in the Fourth Industrial Revolution, which has witnessed the introduction of artificial intelligence in many fields, including health care, psychotherapy, and more (Hounshell, 1984; Mokyr and Strotz, 1998; Brants et al., 2007; Bell, 2019; Thirunavukarasu et al., 2023).

In psychotherapy, the early Eliza program, designed in the 1970s by Weitz Naum, a professor at the Massachusetts Institute of Technology, was a very primitive program, compared to the programs we see today. The program was distinguished by providing some comfort for postgraduate students. Some of them even liked to sit alone next to the computer, and found that the Eliza program helped them a lot, even though they knew it had no emotions, care, or empathy (O'Dell and Dickson, 1984).

On November 22, 2022, ChatGPT-3 became available to the general public. It was a surprise to the technological community and the world, and it was a powerful leap in the field of AI. AI is one of the most advanced areas of modern technology. It was followed by the most famous ChatGPT-4, which is nearly 500 times larger in terms of capacity and also processing capacity. It is the latest version of ChatGPT, launched in March 2023. This is a chatbot that belongs to linguistic artificial intelligence and uses artificial intelligence technology to interact with users in different languages. It has the ability to understand, create, analyze and edit texts, and uses more than 500 billion words from various sources to understand and create texts in smart and creative ways.

Companies then competed to produce large language models in AI: "LLMs." It is an abbreviation of the term "Large Language Models," which refers to AI models that are trained on large amounts of text for the purpose of understanding and generating natural language in an advanced way. Examples include the ChatGPT-3 and 4 from OpenAI, the LaMDA and PaLM models from Google (the basis for Bard), the BLOOM model and XLM-RoBERTa from Hugging Face, and the NeMO model From Nvidia, XLNet, Co:here, and GLM-130B.

Google Bard is a Large Language Model (LLM) created by Google AI. This is a machine-learning model trained on a huge dataset of text and code amounting to 1.56 trillion words. It can generate human-quality text, translate languages, write different types of creative content, and answer questions in a human-like manner. It first appeared on January 18, 2023, when it was announced at the Google AI Conference, and was released to the public on October 16, 2023. Bing AI Chat is a service provided by Microsoft that uses artificial intelligence to improve the search experience of users. Users can interact with Bing as if they were talking to another person, with Bing answering questions and providing information in a natural and friendly way. In addition, Bing can generate images directly from the user's words.

This field has witnessed many important developments in recent years, and it is expected that it will continue to develop in the future at a faster rate and with greater leaps. The AI models allow machines to perform advanced human-like functions. This development began in the 1950s, and continued at varying rates until 2022, when deep learning, a branch of AI, became important in many practical applications such as image recognition and translation (Brants et al., 2007; Bell, 2019; Thirunavukarasu et al., 2023).

The mechanism used in ChatGPT-3 announced by Open AI was a breakthrough that resulted in an artificial intelligence program that can simulate human conversation. Since then, competition has flared among the major companies that had been preparing for such a day for years but were unable to launch a similar produce, namely, Microsoft and Google. Google Barge, Bing, and others introduced large linguistic conversation models that used natural human language relying on a large database; these were trained by interacting with people in specialties and in many fields, including the therapeutic psychological field (Hagendorff and Fabi, 2023; Han et al., 2023).

AI is classified into several categories according to the application, field, and techniques used. In general, it is divided into two types: weak, which is designed to perform a specific task such as voice recognition, and strong, which aims to imitate human intelligence in general (Russell and Norvig, 2010).

This year, large language models have evolved a lot and have reached a stage where they demonstrate human-like language understanding and generation capabilities, which in turn opens new opportunities for using measurement tools to identify the hidden values, attitudes, and beliefs that are encoded in these models. The capabilities of AI to diagnose personality traits and understand feelings and thoughts have been measured and their credibility has been verified by a number of studies (Maksimenko et al., 2018; Kachur et al., 2020; Flint et al., 2022; Han et al., 2023; Landers and Behrend, 2023; Lei et al., 2023; Zhi et al., 2023).

One of the contemporary studies that was concerned with measuring the capabilities of ChatGBT is the study that was presented in the technical report issued by OpenAI on March 27, 2023, in which it conducted tests similar to admission tests in various professional and academic American universities. It included the SATs, the Bar Exam, and the AP final exams. The results showed that the ChatGPT 3.5 and ChatGPT 4.0 are capable of performing human-like on many professional and academic tests.

1.1 Artificial intelligence in psychotherapy field

When a psychologist or counselor carries out the counseling and psychotherapy process, they go through several stages that starting with the preparation phase, which requires several skills, including social intelligence skills. The psychologist employs these skills effectively from the first session and continues until the closing of the sessions. For this reason, previous psychological studies have examined the capabilities of artificial intelligence systems, especially linguistic models, in the therapeutic process. The research is summarized follows:

In the field of diagnosis, artificial intelligence can help improve psychological treatment by providing tools and techniques that help stimulate the process of change and focus on cognitive and emotional understanding (de Mello and de Souza, 2019). It can also contribute to measuring mental (Lei et al., 2023) and emotional disorders and thus reduce the potential risk of suicide (Morales et al., 2017; Landers and Behrend, 2023).

AI can also help improve empirical analysis by developing data-driven models and tools to address new means of selecting therapeutic models (Horn and Weisz, 2020). It can also use speech content analysis and measure mental and emotional disorders as well as the effect of psychiatric medications (Gottschalk, 1999). In addition, AI can use the analysis of physiological signals such as pulse rate, galvanic skin response, and pupil diameter to monitor stress level in users (Zhai et al., 2005).

According to Kachur et al. (2020), AI has ability in the diagnostic process to accurately determine personality traits and has made multidimensional personality profiles more predictable. In another study, Maksimenko et al. (2018) found a relationship between EEG recordings and mental abilities and personality traits. They concluded the importance of designing artificial intelligence programs for personality testing that combine simple tests and EEG measurements to create accurate measurements. Kopp and Krämer (2021) evaluate the ability of intelligent models to visualize and understand mental states speaker and generate behaviors based on them. They concluded that it is necessary to use empathy and positive interactions to support understanding of silent clients.

Regarding the use of smart systems in counseling and psychotherapy, Das et al. (2022) found the effectiveness of GPT2 and DialoGPT in psychotherapy and how the linguistic quality of general conversational models improved through the use of training data related to psychotherapy. Eshghie and Eshghie (2023) showed the ability of ChatGPT to engage in positive conversations, listen, provide affirmations, and introduce coping strategies. Without providing explicit medical advice, the tool was helped therapists make new discoveries.

Likewise, a study of Ayers et al. (2023) evaluated ChatGPT's ability to provide high-quality empathetic responses to patients' questions and found that residents preferred chatbot answers to physician answers. Chatbot responses were rated as more empathetic than doctors' responses. A recent study (Sharan and Romano, 2020) indicated that AI-based methods apply techniques with great efficiency in solving mental health difficulties and alleviating anxiety and depression.

Although previous studies were enthusiastic and tended to support the capabilities of artificial intelligence, there is, in contrast, an opposing view citing errors resulting from AI models in the field of mental health practices. Elyoseph and Levkovich (2023) to compare mental health indicators as estimated by the ChatGPT and mental health professionals in a hypothetical case study focusing on suicide risk assessment. The results indicated that ChatGPT rated the risk of suicide attempts lower than psychologists. Furthermore, ChatGPT rated mental flexibility below scientifically defined standards. These findings have suggested that psychologists who rely on ChatGPT to assess suicide risk may receive an inaccurate assessment that underestimates actual suicide risk.

In addition, research tended to warn against excessive confidence in these systems. Grodniewicz and Hohol (2023) investigate three challenges facing the development of AI systems used in providing psychotherapy services, and explore the possibility of overcoming them: the challenges of deep understanding of psychotherapy strategies, establishing a therapeutic relationship, and the complex voice conversation techniques compatible with humans who convey emotions in their precise structures. The benefits and side effects of using AI in the psychological field should be clarified. Chang et al. (2023) concluded that it is necessary to focus on evaluating the performance of these models, including general performance, response to a task, output, and presentation; their results were heterogeneous in output. Likewise, Woodnutt et al. (2023) found that ChatGPT was able to provide a plan of care that incorporated some principles of dialectical behavioral therapy, but the output had significant errors and limitations, and therefore the potential for harm was possible. Others have pointed out the need to treat AI as a tool but not as a therapist, and limit its role in the conversation to specific functions (Sedlakova and Trachsel, 2023). In addition, there are many challenges that must be overcome before AI becomes able to provide mental health treatment. It is clear that more research is needed to evaluate artificial intelligence to consider how it can be used safely in health care delivery (Grodniewicz and Hohol, 2023). This is why there was an urgent need to conduct this study, which aimed to identify the level of social intelligence of linguistic artificial intelligence models "ChatGPT-4; Bard; Bing" and compare it with psychologists (Bachelor's and Doctorate holders) to reveal the extent to which artificial intelligence contributes to psychotherapy and counseling and to provide comparisons with psychologists.

Consequently, the current study examined the level of social intelligence of artificial intelligence models compared to the performance of psychologists, by using a scale designed to evaluate human social intelligence.

2 Methods

2.1 Participants and procedure

The Human participants were a sample of male psychologists in the Kingdom of Saudi Arabia with one of two levels of education (Bachelor's and doctoral students) at King Khalid University during 2023–2024. The study sample consisted of 180 participants, including 72 bachelor's students and 108 doctoral students in counseling psychological program. They were random selected using stratified method to fit the distribution of participants into two different educational stages. The age of the doctoral students ranged between 33 and 46 years (40.55 ± 6.288), while it was ranged between 20 and 28 years (22.68 ± 7.895) among the bachelor's students.

In this study, a registered version of ChatGPT-4 (OpenAI, 2023) and the free version of Google Bard, and Bing were used. We conducted a single evaluation for each AI model on August 1, 2023 of its SI performance using the Social Intelligence Scale (Sufyan, 1998). In each evaluation, we provided AI the same 64 standard SI scenarios. A link to the questionnaire was sent to human participants via e-mail. While the large linguistic models of AI were asked to answer the scale items individually and their answers were collected in a separate external file by directing a question to the AI models to choose the appropriate answer from the alternative points for each item in the scale.

2.2 Study tools

The performance of the AI models and psychologists was scored using the standard manual (Sufyan, 1998) The SI Scale was prepared by Sufyan (1998) in Arabic to assess SI among adults in similar to the

George Washington University Brief Scale of SI. It consists of 64 items and contained two dimensions: Soundness of judgment of human behavior, which represents the ability to understanding social experiences by observing human behavior. The second dimension assess the ability to act in social situations by analyzing social problems and choosing the best appropriate solutions to them. Sufyan (1998) verified the validity and reliability of this scale. However, the authors of the current study verified the psychometric properties of the scale and its suitability for the objectives of the present study, especially since it will be used to evaluate the performance of large linguistic models on social intelligence skills. Therefore, the scale was presented here to 10 psychology professors at Taiz and King Khalid Universities, and all items were approved, with some items being modified. The modifications of the scale by experts were minor and did not affect the content of the items. Items (1, 7, 12, and 23) were modified grammatically in accordance with the rules of the Arabic language without causing any change in the content of the item.

The validity and reliability sample consisted of 90 individuals from the same research community. Construct validity was verified by examining the correlations between item scores and the total score on the scale using (point, biserial) coefficient. The correlation coefficients ranged between (0.39–0.48) and were significant at the 0.05 level. Construct validity was verified by identifying the significant correlation between the dimensions scores and the total score on the scale using the Pearson correlation coefficient.

The correlation coefficient of the first dimension was 0.82 and in the second dimension, it was 0.73. The reliability of the scale was verified using the re-test method by selecting a sample of 20 undergraduate students from the same research community, and the test was re-tested after 1 month. The reliability coefficient after correction with Spearman's equation was 0.67 for the first dimension and 0.69 for the second dimension, while the overall reliability coefficient was 0.77.

2.3 Scoring

The first dimension's items (41 items) of SI scale were formulated to be answered with true or false (0–1 scores per item; range 0–41), while the answer options of the second dimension (23 items) include 4 points, three of which are false and one is correct (0–1 scores per item; range 0–23).

The total score of SI scale ranged between (0–64), with a higher score indicating higher SI. In all assessments, participants respondents from both human and nonhuman samples were asked to choose the correct answer and the higher the total score, the higher the SI. The SI results of AI models were compared with those of psychologists at both bachelors and doctoral levels.

2.4 Statistical analysis plan

IBM SPSS software (version 28) was used for data analysis. Independent Samples Test was used to examine test–retest reliability of the scale. The relationship between item scores and the total score on the scale was calculated using the point biserial coefficient, while the Pearson correlation coefficient was used to assess the correlation between the dimensions scores and the total score of the scale.

A one-sample *t*-test was used to compare the performance of AI models to the population represented by the psychologists; Means, standard deviations, and percentages were used to determine the ranking of AI models and psychologists.

3 Results

To achieve the research objectives of identifying the level of social intelligence among AI models comparing with psychologists, verification was carried out as follows:

To verify the differences between AI models and psychologists in SI, the average of SI scores for psychologists were extracted; the average scores were 39.19 of bachelor's students and 46.73 of PhD holders. While the raw scores of the AI models were treated as representing independent individual samples (one total score for each model); the scores of SI were 59 of GPT4, 48 of Bing, and 40 of Google Bard.

Therefore, we used a one-sample *t*-test to find out whether these differences were statistically significant, as shown in Table 1.

As per Table 1, the scores of the AI linguistic models are as follows: GPT 4 was 59, Bing was 48, and Google Bard was 40. There are statistically significant differences between ChatGPT-4 and Bing and the psychologists in both academic stages. The AI models have higher SI scores than the psychologists.

As for Google Bard, the result differed; its score was almost equal to that of psychologists with a bachelor's degree, and the differences were not statistically significant. While, its differs compared to doctoral-level, whose average was higher than that of Google Bird in SI. Table 2 shows the level of social intelligence according to the percentile and the raw score for psychologists according to qualification.

The results of this study are summarized as follows:

1 In ChatGPT-4, the score on the SI scale was 59, exceeding 100% of specialists, whether at the doctoral or the bachelor's levels.
2 Bing, whose score on the SI scale was 48, outperformed 50% of doctoral specialists, while 50% of them outperformed him. However, Bing's performance on the SI scale was higher than 90% of bachelor's students.
3 Google Bard, whose score on the SI scale was (40) is superior to only 10% of doctoral holders. Interestingly, 90% of doctoral holders excelled at it. In contrast, Google Bird's performance was higher than 50% of the specialists at the bachelor's level, while 50% of them surpassed it, meaning that Google Bird's performance was equal to the performance of bachelor's students on the SI scale and the differences were not significant.

Figure 1 shows SI levels of AI models and psychologists.

4 Discussion

The main question of this study was "Does artificial intelligence reach the level of human social intelligence?." When we assess humans, we use psychological standards to estimate their level of social intelligence. This is what we did in this study, where the same measure

TABLE 1 The differences between AI and psychologists in the social intelligence.

	Qualification	Mean	Standard deviation	Df	*T*	*p*-value
ChatGPT 59	Bachelor	39.19	7.927	71	21.201	0.00
	Doctoral	46.73	5.974	107	21.341	0.00
Bing 48	Bachelor	39.19	7.927	71	9.426	0.00
	Doctoral	46.73	5.974	107	2.207	0.00
Google Brand 40	Bachelor	39.19	7.927	71	0.862	0.00
	Doctoral	46.73	5.974	107	11.709	0.00

TABLE 2 The level of SI among psychologists according to academic stage.

			Percentages					
		Level	5	10	25	50	75	90
Weighted average (definition 1)	SI	Doctoral	35.90	39.80	44.00	48.00	51.00	54.00
		Bachelor	24.00	25.30	34.25	40.00	46.00	48.70
Tukey's Hinges	SI	Doctoral			44.00	48.00	51.00	
		Bachelor			34.50	40.00	46.00	

was used on the AI represented by the large linguistic model (i.e., ChatGPT 4, Bing, and Google Bard). Our study showed important results regarding the superiority of AI in the field of SI.

The present findings showed that ChatGPT-4 completely outperformed the psychologists. Bing outperformed most of the psychologists at the bachelor's level, while the differences in social intelligence were not significant between Bing and the psychologists at the doctoral level. Interestingly, the psychologists of doctoral holders significantly outperformed Google Bird, while the differences between Google Bird and undergraduate students were not statistically significant, meaning that Google Bird's performance was equal to the performance of bachelor's students on the SI scale.

The result showed that AI outperformed human SI measured by the same scale, and some of it was equal, as in the case of Google Bard, with a certain educational level, which is a bachelor's degree, but it was lower than the level of doctoral. The human participants in this study were a group assumed to have high social intelligence, as many studies have found (Osipow and Walsh, 1973; Wood, 1984), as well as by looking at their average social intelligence measured in the current study compared to the hypothesized mean. By defining social intelligence as the ability to understand the needs, feelings, and thoughts of people in general and to choose wise behavior according to this understanding, it is practically assumed that this would reflected in the superiority of psychologists over the performance of AI. However, our results showed that the differences were of varying, with AI outperforming humans, especially ChatGPT-4, and psychologists with PhDs outperforming Google Bird, while the difference between humans and Ping was not statistically significant.

We believe that the poor performance of Google Bard in SI may be attributed to the date in which this research was conducted, as the Google Bard model was still new and in the early stages of its development, as Google may have been shocked and surprised by what the open AI had achieved. In addition, these results may be due to technical aspects related to the development of the algorithms used in Google Bard. We suggest conducting future studies to track the rapid development of these models, and the extent of their effects on the work of psychotherapists. Another pivotal point that must be pointed out is the ethical extent of the use of artificial intelligence in psychotherapy. Will AI models adhere to the ethics of psychotherapy? Will people want to receive psychotherapy provided by intelligent machines? What about the principles of confidentiality, honesty, empathy, acceptance, and client rights?…etc. These issues need further studies and guidelines for psychotherapists when using artificial intelligence services in counseling and psychotherapy.

What concerns us and those who need counseling and psychotherapy is that this study confirmed the superiority of AI models over humans. These results are partly consistent with the study of Elyoseph and Levkovich (2023) which evaluated the degree of social awareness among the large linguistic models of AI and the extent of the ability of these models to read human feelings and thoughts. They concluded that the ChatGPT was able to provide high-quality responses, and was empathic to patients' questions, with results showing participants' preference for chatbot responses over a doctor's answers. Chatbot responses were also rated as significantly more sympathetic than doctor responses. Some studies that have examined AI for several purposes have indirectly demonstrated the ability of AI in several psychological and mental aspects. Some clients have reported preferring AI-powered assistants over psychotherapists because the assistants were able to deal with their feelings in a distinct and positive manner. It seems like these assistants were able to reflect on the clients' emotions in a way that made them feel comfortable (Ayers et al., 2023; Bodroza et al., 2023; Eshghie and Eshghie, 2023; Haase and Hanel, 2023; Harel and Marron, 2023; Huang et al., 2023).

Another study by Open AI found that GPT4 outperformed humans in postgraduate admission tests in American universities. Literature has indicated that social intelligence is not only an ability in humans but also in artificial intelligence and large linguistic models based on dialog and chat in particular (Herzig et al., 2019). A recent qualitative shift has emerged in the field of artificial intelligence regarding the nature of human intelligence and its effects on the design and development of smart robots. This may create controversy, as social intelligence is added to the behavior of intelligent robots for

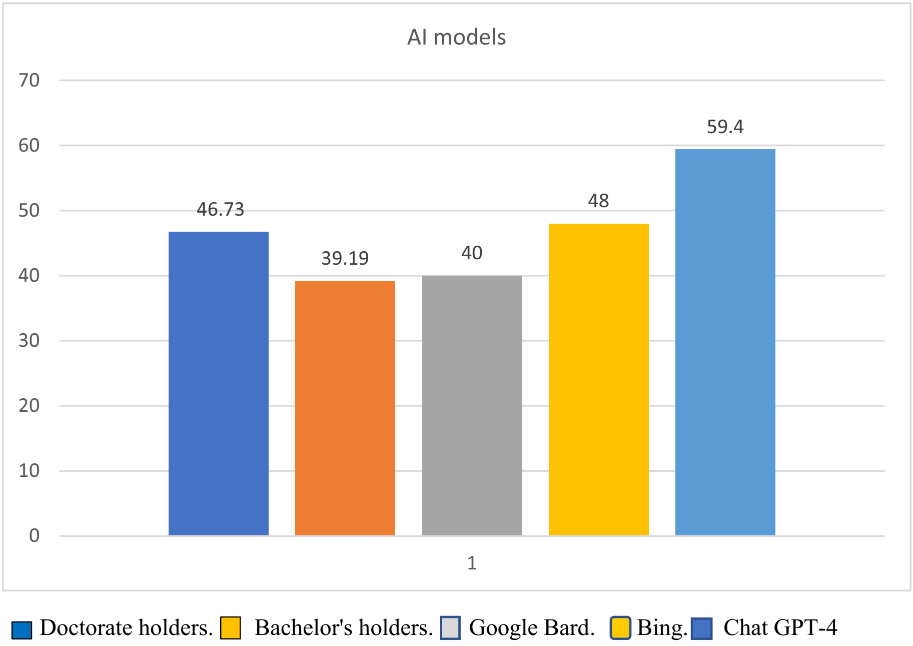


FIGURE 1
Social intelligence levels of AI models and psychologists.

practical purposes and to enable the robot to interact smoothly with other robots or people, that social intelligence may be a stepping-stone toward more human-like artificial intelligence (Dautenhahn, 2007; Guo et al., 2023).

These results confirm the superior ability of AI in SI, as measured by human psychological standards or personality trait tools, and through practical evaluation in conversations conducted between it and clients through the experiments (Herzig et al., 2019; Ayers et al., 2023; Bodroza et al., 2023; Eshghie and Eshghie, 2023; Harel and Marron, 2023).

However, there are references in the literature to concerns and criticisms about AI, some of which relate to errors in diagnoses related to dangerous conditions such as suicide, errors of hallucinations, and fears of moral deviations that need adequate attention and controls in the future studies (Li et al., 2022; Elyoseph and Levkovich, 2023; Grodniewicz and Hohol, 2023). Research also has pointed to a lack of consistency in their responses on psychological measures (Chang et al., 2023), and others have argued that it was necessary to define his role in specific functions (Sedlakova and Trachsel, 2023).

These differences in results may deepen the debate about psychologists' fears of losing their profession to artificial intelligence. Many researchers believe that these fears have accompanied humans during each industrial revolution and ultimately conclude that industrial development helps humans, reduces the less competent individuals, and creates new professions that deal with the new will emerge. Although the changes this time may be more severe, psychologists will not lose their profession, but its form will change in order to adapt to the new developments. The benefit will be much greater than the losses, and the psychologist must absorb the change, live with its rapid development, and contribute to its management.

As for ethical and professional concerns, researchers believe that they are legitimate and realistic concerns, but based on the development of technology throughout history, it is clear that fear accompanies a person for his profession and ethics. However, development continues and it becomes clear that the fears are exaggerated, then some professions or part of them disappear and humans continually adapt to these changes. For example, the printing machine disappeared and there were developments in the secretarial function through the use of computers instead of the printing machine, and cotton workers turned into machine managers. This is why specialists in psychology, psychotherapy and psychiatry recommend absorbing the wave by understanding artificial intelligence and its applications and making the most of this. Developments in counseling and psychotherapy.

Regarding to the ethical aspect, there are legitimate and notable concerns, so we propose multiple forms and sources of solutions to this problem, namely the enactment of laws, the development of algorithms that limit moral deviation during use, and protective programs such as forgery detectors… etc. Since development will pass and will not stop at the limits of our fears, psychotherapists and legislators will need to constantly think about solutions to problems that may affect the profession and its ethics.

In conclusion, the ChatGPT 4 and Bing models have higher social intelligence than psychologists in the bachelor's and doctoral stages, whereas the Bard model is on par with psychologists in the bachelor's category and is outperformed by psychologists in the doctoral stage. According to our results, AI models can be ranked according to their performance on the social intelligence scale from highest to lowest, respectively, as follows: ChatGPT 4, Bing, and finally Google Bard.

The results of the current study can be useful and used to guide psychotherapists in their dealings with clients. Research evaluating the performance of AI models on measures of SI and other aspects of personality is urgently needed to improve the uses of AI in psychotherapy and mental health care planning.

There are some limitations in this study. The sample to verify the psychometric properties of the Social Intelligence Scale was small and homogeneous, and this is a relative shortcoming. This procedure was an additional confirmation since the validity and reliability of the scale had been previously verified by Sufyan (1998). There is a need for future studies that verify validity in a more precise manner on a large sample and in other ways to verify reliability in a more diverse or more precise way.

The social intelligence of the artificial intelligence models was evaluated only once. We were not able to re-evaluate and compare the two evaluations after a period due to the rapid developments in AI applications, which will affect the consistency of results over time. We suggest future longitudinal studies to track changes over time as AI models evolve. We used a subscription version of Chat GPT-4, and free versions of Bing and Google Bird, a difference that may have affected the results given the features available in the paid models compared to the free versions that available to the general public.

It was difficult to obtain a large sample of psychologists in Saudi Arabia, and we relied instead on psychological counseling students at the bachelor's and doctoral levels (there were no master's programs at the time of preparation of the study). We realize that this sample does not represent psychotherapists in the Kingdom of Saudi Arabia. However, it provides a good picture of human performance compared to the performance of AI in the SI scale. On the other hand, the study's sample is confined to male counseling psychology students from a single university. This limited and homogeneous group might not reflect the broader population of psychologists or the general population's social intelligence. Therefore, additional studies with a more diverse and representative sample are needed.

Although the study used a simple and homogeneous sample, its results are an important indicator of the superiority of these industrial systems, even though they appeared a very short time ago as systems simulating human behavior, and it is an indicator of the rapid future development of these systems in the coming years. This study is one of the first studies in this field, as it highlights and documents a historical stage in time for the beginning of the real competition between humans and machines in mental development, and the competition between the systems themselves. The results of the current study is also an indicator of industrial development compared to humans, paving the way for future studies that follow up on these developments and competitions.

Future studies will need to address the limitations of the current study. Our findings provide essential evidence about the degree of social intelligence in AI models that can be evaluated by human standards. These results will have promising future applications in the fields of assessment, diagnosis, and psychotherapy.

It would be fair to point out that the current study evaluated the performance of three different artificial intelligence models and compared them with a reasonable-sized sample of psychologists. In addition, most previous studies did not focus on evaluating social intelligence in artificial intelligence models as much as they focused on evaluating emotional intelligence (for example, Elyoseph et al., 2023), which increases the importance of the current study.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by The Research Ethics Committee at King Khalid University. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

NS: Conceptualization, Data curation, Investigation, Software, Supervision, Writing – original draft. FF: Conceptualization, Methodology, Project administration, Visualization, Writing – review & editing. SA: Conceptualization, Investigation, Visualization, Writing – review & editing. JM: Conceptualization, Data curation, Investigation, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Open Access funding provided by the Qatar National Library.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., et al. (2023). Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* 183, 589–596. doi: 10.1001/jamainternmed.2023.1838

Bell, D. (2019). "The coming of post-industrial society" in *Social stratification, class, race, and gender in sociological perspective. 2nd* ed (New York:Routledge), 805–817.

Bodroza, B., Dinic, B. M., and Bojic, L. (2023). Personality testing of GPT-3: limited temporal reliability, but highlighted social desirability of GPT-3's personality instruments results. *arXiv:2306.04308v2*. doi: 10.48550/arXiv.2306.04308

Brants, T., Popat, A., Xu, P., Och, F. J., and Dean, J. (2007). Large language models in machine translation. In: In Proceedings of the 2007 Joint Conference on Empirical

Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (pp. 858–867).

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., et al. (2023). A survey on evaluation of large language models. *arXiv:2307.03109*. doi: 10.48550/arXiv.2307.03109

Das, A., Selek, S., Warner, A. R., Zuo, X., Hu, Y., Keloth, V. K., et al. (2022). Conversational bots for psychotherapy: a study of generative transformer models using domain-specific dialogues. In: Proceedings of the 21st Workshop on Biomedical Language Processing, 285–297, Dublin: Association for Computational Linguistics.

Dautenhahn, K. (2007). "A paradigm shift in artificial intelligence: why social intelligence matters in the design and development of robots with human-like intelligence" in *50 years of artificial intelligence*. eds. M. Lungarella, F. Iida, J. Bongard and R. Pfeifer, Lecture Notes in Computer Science, vol. *4850* (Berlin, Heidelberg: Springer)

de Mello, F. L., and de Souza, S. A. (2019). Psychotherapy and artificial intelligence: a proposal for alignment. *Front. Psychol.* 10:263. doi: 10.3389/fpsyg.2019.00263

Elyoseph, Z., Hadar-Shoval, D., Asraf, K., and Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Front. Psychol.* 14:1199058. doi: 10.3389/fpsyg.2023.1199058

Elyoseph, Z., and Levkovich, I. (2023). Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. *Front. Psychiatry* 14:1213141. doi: 10.3389/fpsyt.2023.1213141

Eshghie, M., and Eshghie, M. (2023). ChatGPT as a therapist assistant: a suitability study. *arXiv:2304.09873*. doi: 10.48550/arXiv.2304.09873

Flint, S. W., Piotrkowicz, A., and Watts, K. (2022). Use of Artificial Intelligence to understand adults' thoughts and behaviours relating to COVID-19. *Perspect. Public Health*. 142, 167–174. doi: 10.1177/1757913920979332

Gottschalk, L. A. (1999). The application of a computerized measurement of the content analysis of natural language to the assessment of the effects of psychoactive drugs. *Methods Find. Exp. Clin. Pharmacol.* 21, 133–138. doi: 10.1358/mf.1999.21.2.529240

Grodniewicz, J. P., and Hohol, M. (2023). Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. *Front. Psychol.* 14:1190084. doi: 10.3389/fpsyt.2023.1190084

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., et al. (2023). How close is chatgpt to human experts? Comparison corpus, evaluation, and detection. *arXiv:2301.07597*. doi: 10.48550/arXiv.2301.07597

Haase, J., and Hanel, P. H. (2023). Artificial muses: generative artificial intelligence chatbots have risen to human-level creativity. *arXiv:2303.12003*. doi: 10.48550/arXiv.2303.12003

Hagendorff, T., and Fabi, S. (2023). Human-like intuitive behavior and reasoning biases emerged in language models--and disappeared in GPT-4. *arXiv:2306.07622* 3, 833–838. doi: 10.1038/s43588-023-00527-x

Han, N., Li, S., Huang, F., Wen, Y., Su, Y., Li, L., et al. (2023). How social media expression can reveal personality. *Front. Psych.* 14:1052844. doi: 10.3389/fpsyt.2023.1052844

Harel, D., and Marron, A. (2023). Human or machine: reflections on Turing-inspired testing for the everyday. *arXiv:2305.04312*. doi: 10.48550/arXiv.2305.04312

Herzig, A., Lorini, E., and Pearce, D. (2019). Social intelligence. *AI & Soc.* 34:689. doi: 10.1007/s00146-017-0782-8

Horn, R. L., and Weisz, J. R. (2020). Can artificial intelligence improve psychotherapy research and practice? *Admin. Pol. Ment. Health* 47, 852–855. doi: 10.1007/s10488-020-01056-9

Hounshell, D. (1984). *From the American system to mass production, 1800–1932: The development of manufacturing technology in the United States*. Johns Hopkins University Press, Baltimore: JHU Press.

Huang, F., Kwak, H., and An, J. (2023). Is chatgpt better than human annotators? Potential and limitations of chatgpt in explaining implicit hate speech. *arXiv:2302.07736*. doi: 10.1145/3543873.3587368

Kachur, A., Osin, E., Davydov, D., Shutilov, K., and Novokshonov, A. (2020). Assessing the big five personality traits using real-life static facial images. *Sci. Rep.* 10:8487. doi: 10.1038/s41598-020-65358-6

Kopp, S., and Krämer, N. (2021). Revisiting human-agent communication: the importance of joint co-construction and understanding mental states. *Front. Psychol.* 12:580955. doi: 10.3389/fpsyg.2021.580955

Landers, R. N., and Behrend, T. S. (2023). Auditing the AI auditors: a framework for evaluating fairness and bias in high stakes AI predictive models. *Am. Psychol.* 78, 36–49. doi: 10.1037/amp0000972

Lei, L., Li, J., and Li, W. (2023). Assessing the role of artificial intelligence in the mental healthcare of teachers and students. *Soft. Comput.* 1–11. doi: 10.1007/s00500-023-08072-5

Li, X., Li, Y., Liu, L., Bing, L., and Joty, S. (2022). Is gpt-3 a psychopath? Evaluating large language models from a psychological perspective. *arXiv:2212.10529*. doi: 10.48550/arXiv.2212.10529

Maksimenko, V. A., Runnova, A. E., Zhuravlev, M. O., Protasov, P., Kulanin, R., Khramova, M. V., et al. (2018). Human personality reflects spatio-temporal and time-frequency EEG structure. *PLoS ONE* 13:e0197642. doi: 10.1371/journal.pone.0197642

Mokyr, J., and Strotz, R. (1998). The second industrial revolution, 1870–1914. *Stor. dell'Econ. Mond.* 21945, 1–14.

Morales, S., Barros, J., Echávarri, O., García, F., Osses, A., Moya, C., et al. (2017). Acute mental discomfort associated with suicide behavior in a clinical sample of patients with affective disorders: ascertaining critical variables using artificial intelligence tools. *Front. Psych.* 8:7. doi: 10.3389/fpsyt.2017.00007

O'Dell, J. W., and Dickson, J. (1984). Eliza as a "therapeutic" tool. *J. Clin. Psychol.* 40, 942–945. doi: 10.1002/1097-4679(198407)40:4<942::AID-JCLP2270400412>3.0.CO;2-D

OpenAI. (2023). GPT-4 technical report. doi: 10.48550/arXiv.2303.08774

Osipow, S. H., and Walsh, W. B. (1973). Social intelligence and the selection of counselors. *J. Couns. Psychol.* 20, 366–369. doi: 10.1037/h0034793

Russell, S. J., and Norvig, P. (2010). *Artificial intelligence a modern approach*. 3rd Edition, Prentice-Hall, Upper Saddle River: London.

Sedlakova, J., and Trachsel, M. (2023). Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? *Am. J. Bioeth.* 23, 4–13. doi: 10.1080/15265161.2022.2048739

Sharan, N. N., and Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon* 6:e04572. doi: 10.1016/j.heliyon.2020.e04572

Sufyan, N. S. (1998). *Social intelligence and social values and their relationship to psychosocial adjustment among psychology students at Taiz university*. Unpublished doctoral dissertation University of Baghdad, Iraq.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine. *Nat. Med.* 29, 1930–1940. doi: 10.1038/s41591-023-02448-8

Wood, G. B. (1984). The accuracy of counselors' first impressions. Dissertation abstracts international, 45(05), B.

Woodnutt, S., Allen, C., Snowden, J., Flynn, M., Hall, S., Libberton, P., et al. (2023). Could artificial intelligence write mental health nursing care plans? *J. Psychiatr. Ment. Health Nurs.* 31, 79–86. doi: 10.110.1111/jpm.12965

Zhai, J., Barreto, A. B., Chin, C., and Li, C. (2005). User stress detection in human-computer interactions. *Biomed. Sci. Instrum.* 41, 277–282.

Zhi, S., Zhao, W., Wang, R., Li, Y., Wang, X., Liu, S., et al. (2023). Stability of specific personality network features corresponding to openness trait across different adult age periods: a machine learning analysis. *Biochem. Biophys. Res. Commun.* 672, 137–144. doi: 10.1016/j.bbrc.2023.06.012

TYPE Brief Research Report
PUBLISHED 12 March 2024
DOI 10.3389/fpsyg.2024.1348083

Check for updates

OPEN ACCESS

EDITED BY
Simone Belli,
Complutense University of Madrid, Spain

REVIEWED BY
Marcos Novak,
University of California, Santa Barbara,
United States
Dan Zhang,
Tsinghua University, China

*CORRESPONDENCE
Josep M. Fort
✉ josep.maria.fort@upc.edu

RECEIVED 01 December 2023
ACCEPTED 29 February 2024
PUBLISHED 12 March 2024

CITATION
Zhang Z, Fort JM and Giménez Mateu L (2024)
Decoding emotional responses to
AI-generated architectural imagery.
Front. Psychol. 15:1348083.
doi: 10.3389/fpsyg.2024.1348083

Decoding emotional responses to AI-generated architectural imagery

Zhihui Zhang, Josep M. Fort* and Lluis Giménez Mateu

Escola Tècnica Superior d'Arquitectura de Barcelona, Universitat Politècnica de Catalunya, Barcelona, Spain

Introduction: The integration of AI in architectural design represents a significant shift toward creating emotionally resonant spaces. This research investigates AI's ability to evoke specific emotional responses through architectural imagery and examines the impact of professional training on emotional interpretation.

Methods: We utilized Midjourney AI software to generate images based on direct and metaphorical prompts across two architectural settings: home interiors and museum exteriors. A survey was designed to capture participants' emotional responses to these images, employing a scale that rated their immediate emotional reaction. The study involved 789 university students, categorized into architecture majors (Group A) and non-architecture majors (Group B), to explore differences in emotional perception attributable to educational background.

Results: Findings revealed that AI is particularly effective in depicting joy, especially in interior settings. However, it struggles to accurately convey negative emotions, indicating a gap in AI's emotional range. Architecture students exhibited a greater sensitivity to emotional nuances in the images compared to non-architecture students, suggesting that architectural training enhances emotional discernment. Notably, the study observed minimal differences in the perception of emotions between direct and metaphorical prompts among architecture students, indicating a consistent emotional interpretation across prompt types.

Conclusion: AI holds significant promise in creating spaces that resonate on an emotional level, particularly in conveying positive emotions like joy. The study contributes to the understanding of AI's role in architectural design, emphasizing the importance of emotional intelligence in creating spaces that reflect human experiences. Future research should focus on expanding AI's emotional range and further exploring the impact of architectural training on emotional perception.

KEYWORDS

artificial intelligence, emotional perception, architectural imagery, emotional rendering, architectural design, affective computing

1 Introduction

The integration of artificial intelligence (AI) in architectural design marks a significant shift in our engagement with the built environment. This integration challenges traditional perceptions of architecture as a fusion of human emotion and spatial design, a concept echoed by Corbusier and Etchells (2014). The impact of architectural elements such as color, light, and space on human emotions and behaviors, recognized in previous studies (Mehrabian and Russell, 1974; Pallasmaa, 2012; Zhang et al., 2022), underscores the significance of this evolution.

The rise of AI-generated architectural imagery sparks debates within architectural and psychological circles about AI's capacity to evoke emotional resonance akin to human-designed structures (Botros et al., 2023). Public opinion is divided: while some critics argue AI lacks the inherent human touch necessary for genuine emotional engagement (Daniele and Song, 2019; Cetinic and She, 2022; Demmer et al., 2023), others advocate for AI's potential to elicit complex emotional responses (Bagozzi et al., 2022; Cheng et al., 2022). This dichotomy opens up broader inquiries into the role of emotion and perception in AI-enhanced art and design.

Our research delves into the psychological aspects of responses to AI-enhanced architectural imagery. Drawing on interdisciplinary research in human-AI interaction (Ashlock et al., 2023; Zhang et al., 2023), we analyze the emotional reactions of individuals with varying architectural expertise to AI-generated images, including both interior and exterior visualizations. We also investigate the effect of different AI image generation methods on emotional perception (Zhao, 2016). Additionally, the research explores the implications of AI use in architectural education, design practices, and technology evolution, raising philosophical and ethical questions about the interplay between artificial creations and natural human responses.

In conclusion, while acknowledging the limitations of current research, we propose future research directions focused on the synergistic relationship between AI and human designers, and the cultural and social nuances of emotional resonance in AI-generated designs. Our study aims to decode the complex emotional responses triggered by AI in architectural design, contributing to a deeper understanding of behavioral sciences at the intersection of technology and creativity (Pressman, 2001).

2 Literature review

2.1 AI-generated images in architecture

Recent advancements in AI-generated imagery have significantly impacted the intersection of technology and creativity. Göring underscores the capability of AI generators to produce images that are not only highly realistic but also visually appealing, highlighting that the outcome largely depends on the methodology and precision of the text prompts used (Göring et al., 2023). Similarly, Chen delves into the use of deep learning technologies for creating artistic illustrations from concise text descriptions, showcasing AI's ability for style transfer aligned with narrative content, which illustrates the adaptability of AI to various artistic requirements (Chen et al., 2020).

Lu et al. (2024) presents a compelling discovery that humans have a 38.7% success rate in distinguishing real photographs from those generated by AI, suggesting AI's potential to revolutionize visual expression across industries by mimicking reality closely. This could lead to a future where AI not only augments human creativity but also enriches aesthetic environments.

In architecture, Lee et al.'s (2024) research demonstrates AI's capacity to articulate a wide array of design styles in interior spaces, enhancing spatial layouts with specific features, and embodying the design ethos of distinguished architects. Zhang further investigates AI's role as a pivotal tool in architectural design, offering a variety of design solutions and driving innovation. While acknowledging AI's strengths in fostering attractiveness and creativity, Zhang et al. (2023) also notes areas for improvement in authenticity and coherence of the generated designs. Similarly, Akhtar and Ramkumar (2023) views AI more as a collaborator than a substitute in the architectural design process, suggesting that architects can leverage AI to realize innovative solutions and simplify complex tasks.

2.2 Emotion in AI-generated images

The exploration of emotion in AI-generated images, a field emerging at the intersection of affective computing and visual arts, has gained significant momentum. This interdisciplinary area investigates how AI can simulate and evoke human emotional responses through images, a development that reflects the growing sophistication of AI in understanding human emotions (Picard, 2003; Tao and Tan, 2005).

Central to this domain is the capability of AI, particularly machine learning algorithms, to discern and replicate emotional cues in images. These algorithms, trained on extensive emotional datasets, enable AI to generate images that resonate with viewers, paralleling the emotional impact traditionally found in human-created art (Goodfellow et al., 2014; Sun et al., 2022; Gao et al., 2023). Projects like IBM's Watson and OpenAI's CLIP model illustrate AI's potential in creating emotionally engaging visual content (see Figure 1), utilizing advanced techniques to interpret and manipulate emotional content within imagery (Gatys et al., 2016; Radford et al., 2021).

Sentiment analysis, a critical component of affective computing, has been extended to the realm of AI-generated images. This involves algorithms interpreting the emotional tone of images, an approach particularly relevant in the analysis of architectural imagery. The emotional impact of design elements such as spatial composition, color schemes, and textural details can be explored through AI-generated visualizations, offering new insights into architectural design and its emotional resonance (Yildirim, 2022; Enjellina et al., 2023; Ploennigs and Berger, 2023).

However, generating emotional content in images through AI raises significant challenges and ethical considerations. Issues of authenticity in AI-generated emotional expressions and biases in AI-created imagery are major concerns (Zhang et al., 2023). The ethical implications of AI's potential to manipulate emotional responses, particularly in sensitive fields like architectural design, call for careful scrutiny (Chiarella et al., 2022; Futami et al., 2022). The interaction between AI-generated emotions and human responses in architectural imagery is an area of growing interest, with studies focusing on how these images influence human emotional and behavioral responses, and what this implies for the future of architectural experience (Viliunas and Grazuleviciute-Vileniske, 2022; Enjellina et al., 2023).

3 Research hypotheses

In exploring the application of AI in architectural design and its impact on emotional perception, this study aims to

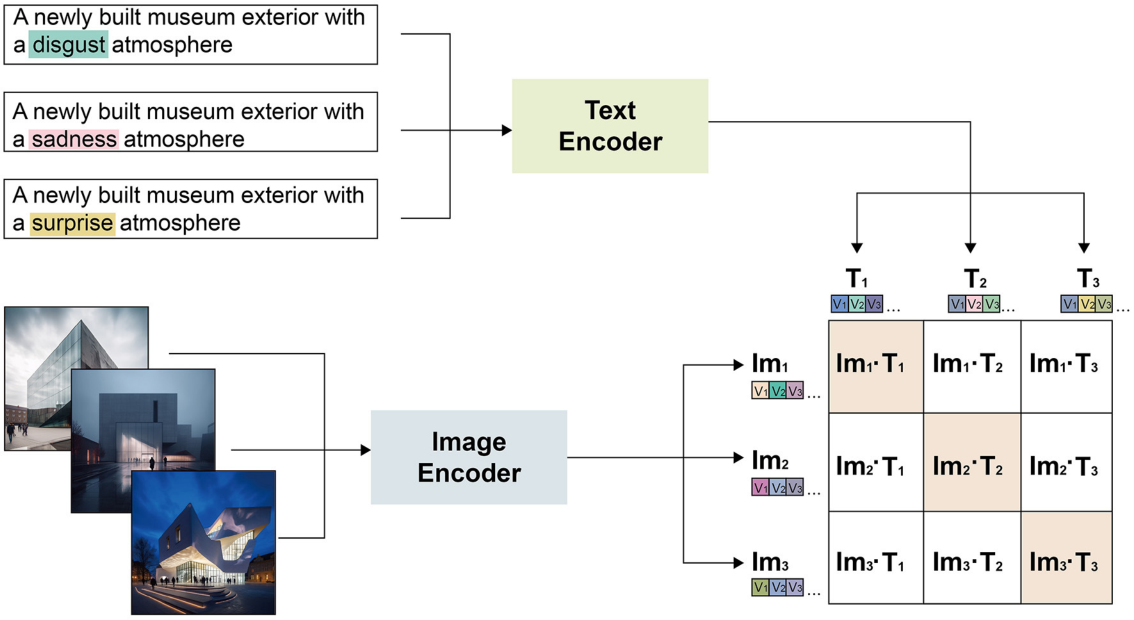


FIGURE 1
Technological advancements in emotionally resonant image generation: principles of techniques similar to CLIP for embedding emotional context into images.

validate the following hypotheses, which are formulated based on prior research and theoretical frameworks. These hypotheses serve as the foundation for the study's design and methodology, guiding our investigation into the emotional role of AI in architectural visual representation and its impact across different audience groups.

Hypothesis 1 (H1): AI-generated architectural images are capable of effectively eliciting specific emotional responses, demonstrating similar or superior emotive resonance compared to human-designed structures.

Hypothesis 2 (H2): There is a significant difference in emotional perception of AI-generated architectural images between architecture students (Group A) and non-architecture students (Group B). This difference is attributed to the professional training of architecture students, making them more sensitive to the emotional details in the images.

Hypothesis 3 (H3): AI-generated architectural images created within interior settings, such as homes, are more effective in expressing emotions compared to those generated in exterior settings, such as museums.

By systematically validating these hypotheses, we aim to contribute to the ongoing discourse on the integration of AI in architectural design, particularly in terms of enhancing emotional engagement and understanding among diverse groups.

4 Method

4.1 Software and tools selection

The choice of software and tools was crucial in our research aimed at examining how various AI technologies render emotional content in architectural imagery. We conducted an evaluation of five prominent image generation software: Stable Diffusion (Version 1.5 with LDMs Algorithms) (Pinaya et al., 2022), DeepFloyd IF (Stability, 2023), DALL E2 (Open, 2023), Midjourney (Version 5.1) (Midjourney, 2023), and Photoshop 2023 (Adobe, 2023). Each tool was tested using two prompts designed to generate images of a newly built museum exterior, with one prompt emphasizing a "happy atmosphere" and the other focusing on creating a "cheerful atmosphere that reflects happiness." This approach allowed us to generate a collection of images for a comparative analysis of each software's ability to capture and convey the emotional essence of the prompts.

In our analysis, Midjourney distinguished itself by most accurately reflecting the intended emotional tones of the prompts. DeepFloyd IF demonstrated a somewhat limited correlation with the specified emotional content. Other tools, including Stable Diffusion, DALL E2, and Photoshop 2023, showed varying degrees of effectiveness in recognizing and rendering the emotional subtleties embedded in the prompts. Open-source platforms like Stable Diffusion and DeepFloyd IF offer extensive customization through plugins like ControlNet, providing detailed control over image generation aspects. The potential integration of technologies like Dreambooth and loRA with these platforms hints at future advancements in developing emotion-specific AI models. Conversely, DALL E2 and Photoshop 2023, while excelling in localized and extensive image modifications, did not align as effectively with our specific research focus on emotional expression in AI-generated architectural visuals.

We selected Midjourney as our primary tool, primarily due to its proficiency in generating images that resonated emotionally from text descriptions. This choice underscores our research intent to delve into AI's capability to evoke specific emotional responses through architectural imagery, a vital component in understanding the nuances of human-AI interaction within behavioral sciences.

4.2 Artificial intelligence in architectural rendering

Our research utilized the Midjourney AI software for generating images, focusing on two architectural settings: "home interior" and "museum exterior." The choice of a "home interior" setting was driven by its universal relevance in daily life, providing a familiar context for eliciting and analyzing emotional responses. On the other hand, the "museum exterior" was selected for its cultural and public significance, offering a diverse spectrum of emotional engagement possibilities.

The crafting of prompts for AI image generation was a key element in our study. We aimed to explore how AI interprets and visualizes emotions within architectural contexts. To achieve this, we developed two types of prompts: one incorporating explicit emotional descriptors, such as "joy," and another utilizing metaphorical language to convey emotions, like "creates a cheerful atmosphere that reflects happiness." This dual approach allowed us to assess the effectiveness of both direct and metaphorical expressions in translating emotions into AI-generated architectural images.

Informed by Ekman's (2005) theory of basic emotions, our study encompassed six emotions: joy, sadness, anger, fear, surprise, and disgust. This range was integral to examining a wide array of emotional responses in architectural environments. For each emotion, we generated images for both the "home interior" and "museum exterior," culminating in a diverse set of 24 architectural images (see Figure 2).

For example, to create imagery for a "museum exterior," we employed prompts like "A newly built museum exterior with a happy atmosphere" to directly express the emotion and "a newly built museum exterior creates a cheerful atmosphere that reflects happiness" for a metaphorical representation. This method was consistently applied across different emotions and replicated for the "home interior" settings. The variation in emotional content of the prompts was crucial in our exploration of how AI-generated architectural renderings could mirror and potentially influence human emotions, a topic of great significance in behavioral sciences.

4.3 Survey design

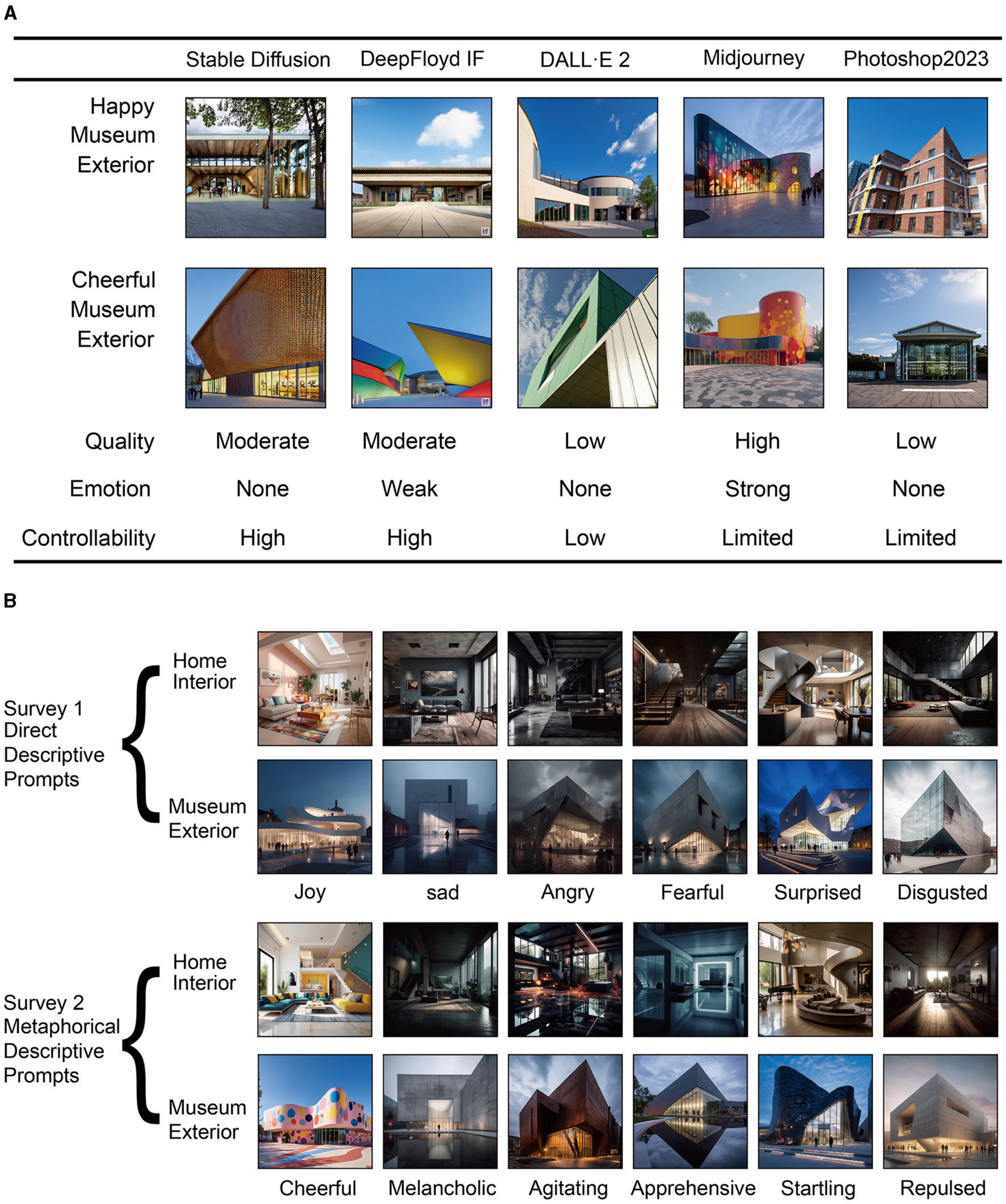
The survey's design was a crucial element in our investigation into the emotional responses elicited by AI-generated architectural images. Our study aimed to discern the impact of these images on both architectural professionals and the general public, focusing on the psychological aspects of their reactions.

To optimize participant engagement and reduce fatigue, the survey was structured into two separate sections. Each section presented participants with a series of AI-generated architectural images. These images were aligned with specific emotions, conveyed either through direct or metaphorical language. Participants were instructed to rate their immediate emotional response to each image using a 0 to +10 scale. On this scale, 0 represented a very weak emotional response, while 10 denoted a highly intense reaction. This rating method was devised to encompass a broad spectrum of emotions, including happiness, sadness, anger, fear, surprise, and disgust. We encouraged participants to trust their gut reactions, underlining the subjective nature of the survey and affirming that there were no right or wrong answers. In designing our survey, we opted for a framework that participants could easily understand and engage with to assess the emotional responses elicited by AI-generated architectural imagery. Therefore, we employed the six basic emotions framework due to its clarity and ease of explanation to participants. Some more recent emotional theory frameworks, such as those proposed by Cowen et al. (2020) and Tang et al. (2023), offer a broader range of emotions and dimensions that, while providing detailed insights into emotional experiences, could potentially confuse participants in this study and significantly increase the workload involved in conducting the survey and analyzing the data. Therefore, we did not adopt these more complex emotional frameworks.

Furthermore, we decided against using a binary approach to emotional analysis, such as the positive and negative polarity, due to its limited capability in capturing the rich emotional engagement we aimed to explore with architectural imagery. In our previous experiments on emotional assessment, including the use of the Self-Assessment Manikin (SAM) questionnaire, participants indicated challenges in comprehension and the need for extensive explanation, impacting the efficiency and effectiveness of the survey (Zhang et al., 2022). Feedback from these preliminary trials revealed a preference among participants for the basic emotions model, which they found to be more intuitive and relatable. By choosing the six basic emotions, our study aimed to maintain a clear and consistent evaluative framework, effectively capturing the subtleties of people's emotional responses to AI-generated architectural imagery without the complexities and ambiguities associated with more elaborate emotional frameworks. Overall, the survey was meticulously designed to probe the intricate relationship between AI-generated images and human emotions, particularly in the context of architectural visualization. This methodology was central to our overarching goal of uncovering and understanding emotional reactions within architectural settings, thereby enriching the discourse in behavioral sciences.

4.4 Participants

In our study, the selection of participants was crucial for exploring the emotional and psychological responses to AI-generated architectural imagery. Inspired by Garip and Garip's (2012) findings, which indicate aesthetic differences between architecture and non-architecture students, we sought to investigate how such disparities might influence the perception and emotional response to AI-enhanced architectural visuals. To this end, we recruited 789 university students, aged 19–40, and divided them into two distinct groups: architecture majors (Group A, comprising 389 participants) and non-architecture majors (Group B, comprising 400 participants). This division was strategically chosen to assess the impact of educational and professional backgrounds on the engagement with AI-generated architectural imagery, grounding our



A

	Stable Diffusion	DeepFloyd IF	DALL·E 2	Midjourney	Photoshop2023
Happy Museum Exterior					
Cheerful Museum Exterior					
Quality	Moderate	Moderate	Low	High	Low
Emotion	None	Weak	None	Strong	None
Controllability	High	High	Low	Limited	Limited

FIGURE 2
(A) Comparative analysis of image generation software, **(B)** architectural images generated through AI, illustrating the visual expression of the six basic emotions examined in the experiment.

participant selection in the premise that professional training and educational experiences significantly shape aesthetic judgment and emotional interactions with architectural design.

Group A, consisting of architecture students, was presumed to possess a deeper understanding and critical appreciation of architectural design. This expertise was expected to influence their emotional responses, with a potential focus on technical and aesthetic aspects of the AI-generated images.

In contrast, Group B included students from diverse non-architecture disciplines, representing a broader demographic akin to the general public. Their reactions were hypothesized to be more rooted in instinctual emotional responses, offering insights into how AI-generated architectural visuals are perceived by those outside the architectural field.

In adherence to the Sex and Gender Equity in Research (SAGER) guidelines, our study consciously did not collect gender-specific data, aiming to eliminate potential gender bias. This decision was made to ensure that our findings were focused solely on emotional and perceptual responses, irrespective of gender (Heidari et al., 2016).

The comparative analysis between these two groups was designed to provide a holistic understanding of how different educational backgrounds affect the perception and emotional engagement with AI-generated architectural imagery. Insights gained from this study are expected to contribute significantly to the fields of behavioral sciences and architectural design, particularly in understanding how AI-generated visuals are received and interpreted by diverse audiences.

4.5 Analysis strategy

The strategy for analyzing our data revolves around three core pillars, each designed to thoroughly investigate the role of AI in creating emotionally resonant architectural imagery. This tripartite approach allows us to delve into both the emotive capacity of AI-generated images and the perceptual differences in their reception among varied audiences.

4.5.1 Assessment of emotive expressivity in AI-generated images

The first aspect of our analysis is dedicated to evaluating the emotional expressiveness of AI-generated architectural images. This involves examining how closely the emotions conveyed in the AI-generated prompts align with the emotions perceived by participants. By assessing this alignment, we aim to understand the effectiveness of AI in accurately interpreting and rendering the intended emotional content within architectural visualizations. This analysis is crucial in uncovering the psychological impact these AI-generated images have on viewers.

4.5.2 Effectiveness of descriptive methods in emotional conveyance

Our second pillar concentrates on comparing the effectiveness of two descriptive approaches “direct descriptive words vs. metaphorical language” in AI-generated images. The goal here is to determine which method more effectively communicates the intended emotional context within the imagery. This comparison is vital for understanding the influence of language in shaping emotional perception in AI-generated architectural visuals.

4.5.3 Differential interpretation between professional and lay audiences

The third pillar of our analysis distinguishes the perceptual differences between architectural professionals and the general public in response to AI-generated architectural images. This comparison seeks to gauge the utility of AI imagery as a tool for professional use in architecture, as well as its role in facilitating intricate and nuanced architectural representations. Analyzing the variances in emotional and perceptual responses between these groups offers insights into how AI-generated imagery is interpreted across diverse audiences.

In order to ensure a consistent rating scale across all images, we normalized the original average scores for each emotion. The normalized score rate P_{ij} for emotion j on image i is calculated as follows(refer to Equation 1):

$$P_{ij} = \frac{E_{ij}}{S_i \times n} \tag{1}$$

where, E_{ij} is the original average score for emotion j on image i; S_i is the sum of the average scores for all emotions on image i; n is the number of images, which is six in our study.

This normalization process ensures that the sum of the scores for all emotions on each image equals 1/6, allowing for a fair comparison of the relative prominence of each emotion across different images. The normalized score rate P_{ij} reflects the proportion of the average score for emotion j relative to the average scores for all emotions on the given image.

In the analysis of data across different groups within our study, we adapted our statistical approach based on the specific characteristics of the dataset. For datasets exhibiting a normal distribution, the analysis was conducted using t-tests to compare means, alongside the calculation of Cohen’s d to provide a measure of effect size, following the guidelines set by Schmidt and Bohannon (1988). In instances where the dataset deviated from normal distribution, the Wilcoxon signed-rank test was employed as a non-parametric alternative, with Cliff’s (1993) delta utilized to assess the magnitude of the observed effects.

To facilitate our data analysis process, we leveraged a suite of Python libraries tailored for statistical computing and visualization. This included the use of NumPy for handling array-based numerical computations, Pandas for its powerful data structure and analysis tools, SciPy for conducting both parametric and non-parametric statistical tests, and plotly for creating visual representations of our findings. The integration of these tools not only bolstered the thoroughness of our statistical examination but also enhanced the clarity and interpretability of the results presented.

5 Result

5.1 Emotional expression across all groups

Our comprehensive analysis of AI-generated architectural images across all participant groups revealed notable trends in emotional expression, which is illustrated in Figure 3.

5.1.1 Direct prompt (home and museum settings)

In the direct prompt category for home settings, joy emerged as the most prevalently expressed emotion at 65.87%. The least effectively conveyed emotion was anger, registering only 11.02%. Other emotions like sadness, fear, surprise, and disgust ranged between 16.53 and 22.92%. In the museum settings under direct prompts, joy still led at 42.63%, but with a notable reduction

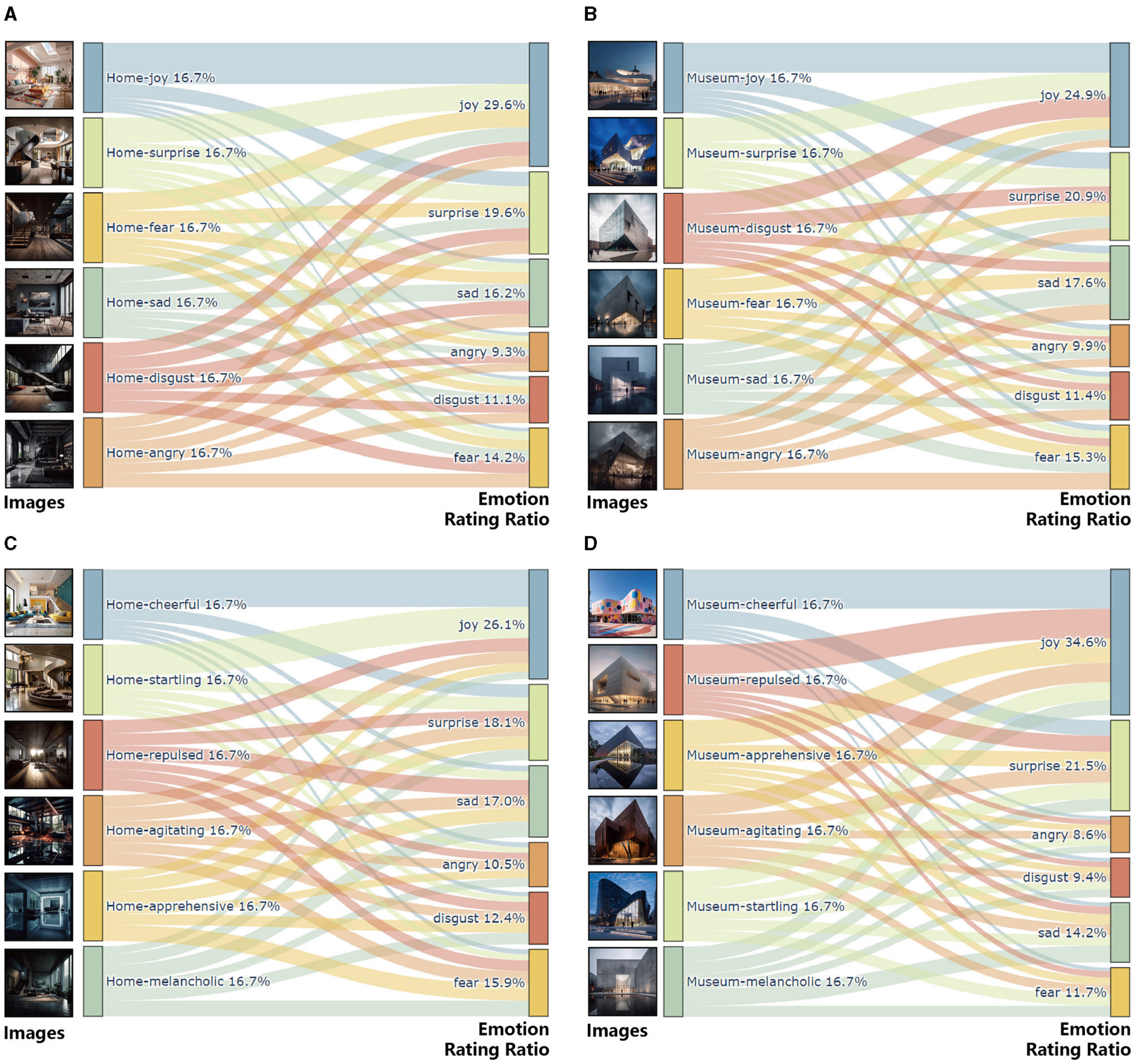


FIGURE 3
(A) Sankey diagram illustrating emotion ratings distribution for home images generated by direct prompts, **(B)** Sankey diagram illustrating emotion ratings distribution for museum images generated by direct prompts, **(C)** Sankey diagram illustrating emotion ratings distribution for home images generated by metaphoric prompts, **(D)** Sankey diagram illustrating emotion ratings distribution for museum images generated by metaphoric prompts.

in effectiveness compared to home settings. Disgust recorded the lowest effectiveness at 10.60%.

5.1.2 Metaphoric prompt (home and museum settings)

With metaphoric prompts, joy remained the dominant emotion in home settings, scoring 57.19%. The lowest effectiveness was again seen in anger at 11.98%. For museum settings, joy's effectiveness slightly decreased to 56.05%, with disgust being the least effective at 7.72%.

5.1.3 Patterns in highest-rated emotions

Our analysis of the highest-rated emotion for each image revealed some intriguing patterns:

- In home settings with direct prompts, joy dominated other emotions, even when the prompts were intended to evoke different emotions like sadness or fear.
- In museum settings, the results were more mixed, with joy still prevailing but to a lesser extent, indicating potential ambiguities in emotional expression.

5.1.4 Positive and negative emotional performance

When categorizing emotions as positive and negative, we observed:

- In home settings with direct prompts, positive emotions like joy, surprise, and disgust outperformed negative emotions like sadness and fear.
- In museum settings, the performance gap between positive and negative emotions was narrower, with anger and disgust showing reduced effectiveness.

5.1.5 Indoor vs. outdoor image analysis

The study also indicated that indoor images generally conveyed emotions more effectively than outdoor images. This trend was consistent across both direct and metaphorical prompts, suggesting that the spatial context significantly influences emotional perception in AI-generated imagery.

In summary, our results indicate a trend where AI-generated images are more effective in conveying positive emotions, particularly joy, across different settings and prompt types. The effectiveness of emotional expression also appears to be influenced by the architectural context, with indoor images demonstrating a higher capacity for emotional conveyance. These findings offer significant insights into the capabilities and limitations of AI in architectural visualization, particularly in its ability to resonate emotionally with viewers from diverse backgrounds.

5.2 Emotional expression across Group A and Group B

Our study's analysis of AI-generated architectural images revealed distinct patterns in emotional perception between architecture students (Group A) and non-architecture students (Group B), as outlined in Figure 4.

5.2.1 Comparison between direct and metaphoric prompts

Within Group A, we observed minimal differences in the perception of surprise and sadness between direct and metaphorical prompts. This suggests a uniformity in emotional interpretation regardless of the prompt type for these emotions. Group B, however, showed a more pronounced difference in the perception of sadness across the prompt types, indicating varied emotional interpretations based on educational background.

5.2.2 Direct prompt analysis

Comparing the emotional perception under direct prompts between the two groups, we noted significant differences across all emotions. The variance was particularly notable for joy and surprise (*P*-value of 0.0235), suggesting these emotions are universally perceived but with subtle differences influenced by the viewer's background. Metaphorical prompts also demonstrated differences in emotional scores between the groups, with the least variance in fear and sadness (*P*-value of 0.0236).

5.2.3 Individual emotional expression

In Group A's direct prompt for home settings, joy was the most prominently expressed emotion at 66.01% Group B showed a similar trend, with joy being the most effectively expressed emotion. However, the overall performance of emotional expression in Group B was slightly lower than in Group A.

5.2.4 Analysis of highest-rated emotion

In Group A's direct prompt (home), joy dominated the ratings even for images intended to convey other emotions like sadness or fear. The museum settings in Group A showed a mixed pattern, with joy still leading but with less dominance compared to home settings.

5.2.5 Positive and negative emotional performance

The binary analysis revealed that positive emotions, particularly joy, were better conveyed in AI-generated images than negative emotions across both groups. The effectiveness varied depending on the setting, with indoor (Home) images generally showing better emotional rendering ability than outdoor (Museum) images. Overall, our results indicate significant differences in the emotional perception of AI-generated architectural images between architecture and non-architecture students. These findings provide valuable insights into how educational background influences emotional interpretation of AI-generated images, with implications for the utilization of AI in architectural design and its perception by different audience segments.

6 Discussion

This study provides an in-depth comparison of the emotional perceptions of two distinct groups toward AI-generated architectural images, unveiling several key findings and their broader implications.

6.1 The technical challenges of emotional expression

Our research delves into the significant challenges AI faces in encoding complex emotions into visual forms. While "joy" has been consistently and effectively depicted, our findings show that other emotions such as anger, sadness, and fear are less accurately portrayed. This discrepancy underscores the inherent difficulty of translating the nuanced spectrum of human emotions into AI-generated imagery. It highlights the imperative need for the development of more advanced AI models that can more finely understand and reflect the complexity of human emotions. For instance, the exploration of deep learning and neural networks presents a promising avenue to enhance AI's capability in emotion

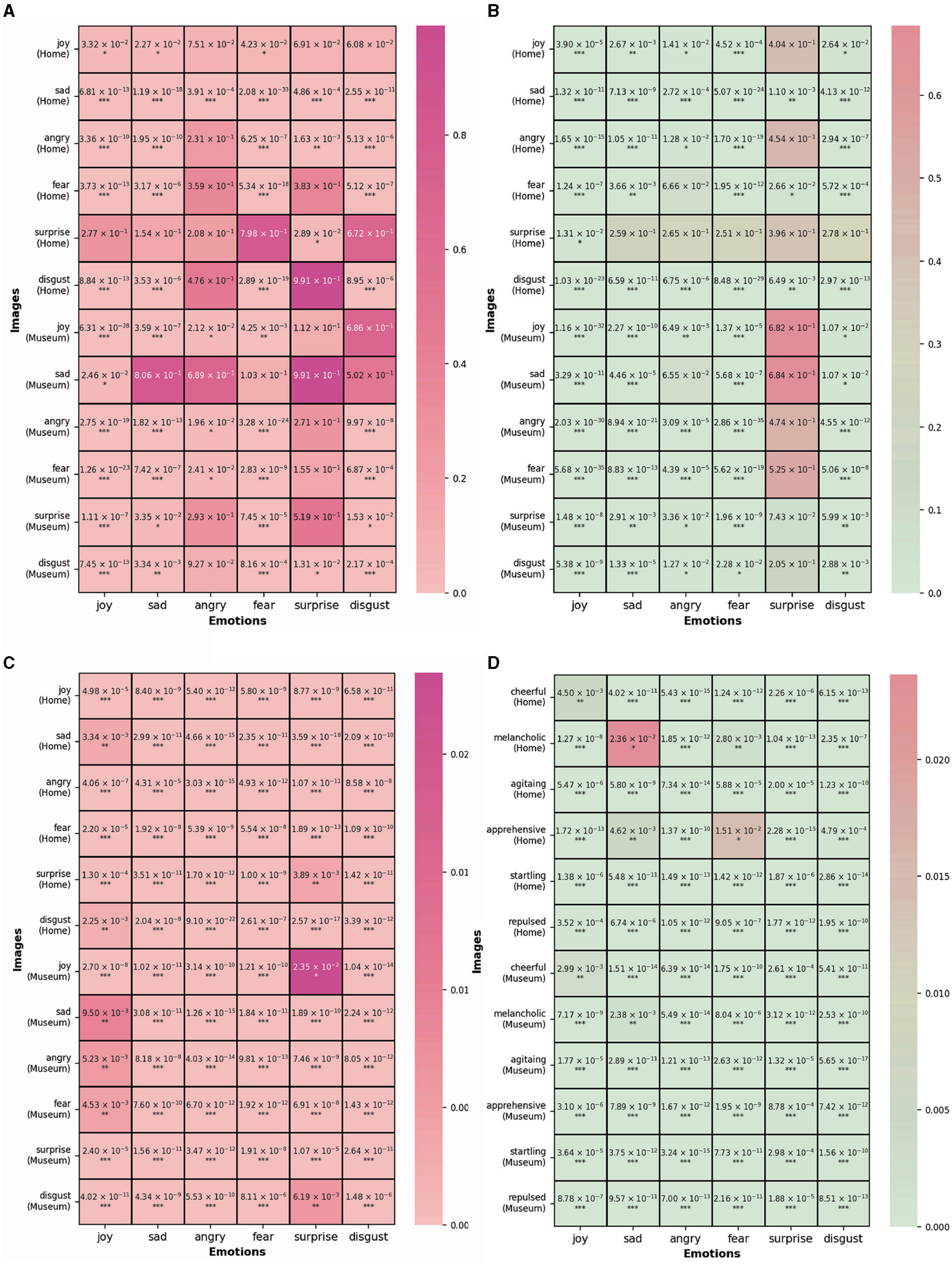


FIGURE 4
(A) Comparative analysis of emotion rating for Group A—heatmap of direct prompt vs. metaphoric prompt, **(B)** comparative analysis of emotion rating for Group B—heatmap of direct prompt vs. metaphoric prompt, **(C)** comparative analysis of emotion rating between Group A and Group B using direct prompt, **(D)** comparative analysis of emotion rating between Group A and Group B using metaphoric prompt. **P*-value < 0.05; ***P*-value < 0.01; ****P*-value < 0.001.

understanding and expression. These findings suggest a divergence from our initial hypotheses, indicating that while AI shows potential in emotional representation, its current abilities to capture and convey the full range of human emotions are limited.

6.2 The impact of educational background on emotional interpretation

Our research revealed significant differences in emotional perception between architecture students (Group A) and non-architecture students (Group B), aligning with our hypothesis. These differences can likely be attributed to the specialized training of architecture students, who are educated to understand the interplay between spatial design and emotional evocation. This finding highlights how educational and professional training shapes individuals' emotional interpretation of architectural spaces. It underscores the importance of interdisciplinary collaboration, integrating AI technology and emotional understanding in the educational process to provide designers and architects with a more comprehensive training. This synergy between AI and architectural education not only validates our initial hypothesis but also opens up new avenues for enriching the emotional depth of architectural design through AI.

6.3 The environmental impact on emotional rendering

Our study emphasizes the critical role of architectural imagery in eliciting emotional responses, with a notable finding that AI-generated images of indoor settings, such as homes, are more effective in emotional rendering than those of outdoor settings like museums. This distinction between indoor and outdoor environments in terms of emotional expression aligns with our hypothesis and is crucial for understanding the application of AI in architectural design. Emotional rendering, alongside aesthetic and functional considerations, plays a vital role in architectural imagery.

Expanding upon the differences in emotional expression between indoor and outdoor environments, our analysis delves into how specific features of these settings influence the conveyance and perception of emotions. The findings suggest that future research should employ a broader array of environmental samples to validate and further explore these insights. Designers can leverage this knowledge to optimize spatial design, enhancing emotional resonance within architectural spaces.

While these observations are consistent with our initial hypotheses, it is important to acknowledge the limitations of our experimental setup, particularly the range of environments we were able to include. A more extensive exploration of different settings is necessary to deepen our understanding of how environmental factors impact emotional responses. This future research direction could offer more nuanced insights into the complex interplay between AI-generated architectural imagery and human emotion.

6.4 Limitations and future directions

Despite the valuable insights provided, our study faces limitations due to the rapid evolution of AI image generation software and the selected environmental settings, which may not fully represent the diverse architectural contexts. The reliance on a limited number of prompts to explore emotional conveyance and the lack of detailed analysis on the influence of architectural training underscore areas for future investigation.

Future studies should expand the variety of prompts and environments to capture a broader spectrum of emotional responses. A more detailed examination of the impact of architectural education on emotional perception could offer deeper insights, considering factors such as the duration and specificity of training experiences.

Moreover, the binary approach to emotional analysis in our study simplifies the complex nature of human emotions. Future research should employ more nuanced methods to analyze the multidimensional aspects of emotional responses, possibly incorporating multisensory elements beyond the visual to enrich the understanding of architectural imagery's emotional impact.

By addressing these limitations, subsequent research can enhance our comprehension of AI's role in architectural design, potentially leading to the development of practices that resonate more profoundly on an emotional level with diverse audiences.

7 Conclusion

This study sheds light on the nuanced capabilities and limitations of AI in evoking emotions within architectural imagery, revealing AI's proficiency in depicting joy and its superior emotional rendering in indoor environments. Particularly, architecture students displayed enhanced sensitivity to AI-generated images, likely due to their specialized training. These findings underscore AI's potential in bridging technological innovation with human emotional experiences in architectural design, suggesting a future where AI not only enhances aesthetic appeal but also fosters emotionally resonant spaces. This research marks a significant step toward understanding AI's role in architecture, emphasizing the importance of integrating emotional intelligence in design practices to create spaces that resonate with human experiences.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: 10.6084/m9.figshare.23896818.

Ethics statement

The studies involving humans were approved by Ethics Committee of the Universitat Politècnica de Catalunya. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. No animal studies are presented in this manuscript.

Author contributions

ZZ: Writing – original draft, Visualization, Supervision, Software, Methodology, Formal analysis, Data curation, Conceptualization. JF: Writing – review & editing, Supervision, Conceptualization. LG: Writing – review & editing, Methodology, Data curation, Conceptualization.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adobe (2023). Photoshop Desktop (September 2023 Release). Available online at: https://www.adobe.com/es/products/photoshop

Akhtar, M. H., and Ramkumar, J. (2023). "AI in architecture: architects do not like AI. Is it?" in *AI for Designers* (Berlin: Springer), 67–84. doi: 10.1007/978-981-99-6897-8_4

Ashlock, D., Maghsudi, S., Liebana, D. P., Spronck, P., and Eberhardinger, M. (2023). *Human-Game AI Interaction: Report from Dagstuhl Seminar 22251.*

Bagozzi, R. P., Brady, M. K., and Huang, M. H. (2022). AI service and emotion. *J. Serv. Res.* 25, 499–504. doi: 10.1177/10946705221118579

Botros, C. R., Mansour, Y., and Eleraky, A. (2023). Architecture aesthetics evaluation methodologies of humans and artificial intelligence. *MSA Eng. J.* 2, 450–462. doi: 10.21608/msaeng.2023.291897

Cetinic, E., and She, J. (2022). "Understanding and creating art with AI: review and outlook," in *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (New York, NY: ACM), 18. doi: 10.1145/3475799

Chen, Z., Chen, L., Zhao, Z., and Wang, Y. (2020). "AI illustrator: art illustration generation based on generative adversarial network," in *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)* (Beijing: IEEE), 155–159. doi: 10.1109/ICIVC50857.2020.9177494

Cheng, K.-T., Chang, K., and Tai, H.-W. (2022). AI boosts performance but affects employee emotions. *Inf. Resour. Manag. J.* 35, 1–18. doi: 10.4018/irmj.314220

Chiarella, S. G., Torromino, G., Gagliardi, D. M., Rossi, D., Babiloni, F., Cartocci, G., et al. (2022). Investigating the negative bias towards artificial intelligence: effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings. *Comput. Hum. Behav.* 137:107406. doi: 10.1016/j.chb.2022.107406

Cliff, N. (1993). Dominance statistics: ordinal analyses to answer ordinal questions. *Psychol. Bull.* 114, 494–509. doi: 10.1037/0033-2909.114.3.494

Corbusier, L., and Etchells, F. (2014). *Towards a New Architecture (Reprint or 1927 Edition)*, New York, NY: Dover Publications, 289.

Cowen, A. S., Fang, X., Sauter, D., and Keltner, D. (2020). What music makes us feel: at least 13 dimensions organize subjective experiences associated with music across different cultures. *Proc. Natl. Acad. Sci.* 117, 1924–1934. doi: 10.1073/pnas.1910704117

Daniele, A., and Song, Y. Z. (2019). "AI + art = human," in *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY: ACM), 155–161. doi: 10.1145/3306618.3314233

Demmer, T. R., Kühnapfel, C., Fingerhut, J., and Pelowski, M. (2023). Does an emotional connection to art really require a human artist? Emotion and intentionality responses to AI- versus human-created art and impact on aesthetic experience. *Comput. Hum. Behav.* 148:107875. doi: 10.1016/j.chb.2023.107875

Ekman, P. (2005). "Basic emotions," in *Handbook of Cognition and Emotion*, eds T. Dalgleish, and M. J. Power (Hoboken, NJ: John Wiley & Sons), 45–60. doi: 10.1002/0470013494.ch3

Enjellina, Beyan, E. V. P., and Rossy, A. G. C. (2023). A review of AI image generator: influences, challenges, and future prospects for architectural field. *J. ArtifIntell. Archit.* 2, 53–65. doi: 10.24002/jarina.v2i1.6662

Futami, K., Yanase, S., Murao, K., and Terada, T. (2022). Unconscious other's impression changer: a method to manipulate cognitive biases that subtly change others' impressions positively/negatively by making AI bias in emotion estimation AI. *Sensors* 22:9961. doi: 10.3390/s22249961

Gao, T., Zhang, D., Hua, G., Qiao, Y., and Zhou, H. (2023). "Artificial intelligence painting interactive experience discovers possibilities for emotional healing in the post-pandemic era," in *International Conference on Human-Computer Interaction* (Berlin: Springer), 415–425. doi: 10.1007/978-3-031-35998-9_56

Garip, E., and Garip, B. (2012). Aesthetic evaluation differences between two interrelated disciplines: a comparative study on architecture and civil engineering students. *Procedia-Soc. Behav. Sci.* 51, 533–540. doi: 10.1016/j.sbspro.2012.08.202

Gatys, L., Ecker, A., and Bethge, M. (2016). A neural algorithm of artistic style. *J. Vis.* 16, 326–326. doi: 10.1167/16.12.326

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 27.

Göring, S., Rao, R. R. R., Merten, R., and Raake, A. (2023). Analysis of appeal for realistic AI-generated photos. *IEEE Access* 11, 38999–39012. doi: 10.1109/ACCESS.2023.3267968

Heidari, S., Babor, T. F., Castro, P. D., Tort, S., and Curno, M. (2016). Sex and gender equity in research: rationale for the sager guidelines and recommended use. *Res. Integr. Peer Rev.* 1, 1–9. doi: 10.1186/s41073-016-0007-6

Lee, J.-K., Jeong, H., Kim, Y., Choi, S., Jo, H., Chae, S., et al. (2024). "How to enhance architectural visualisation using image gen AI," in *Multimodality in Architecture: Collaboration, Technology and Education* (Berlin: Springer), 157–173. doi: 10.1007/978-3-031-49511-3_9

Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., et al. (2024). Seeing is not always believing: benchmarking human and model perception of AI-generated images. *Adv. Neural Inf. Process. Syst.* 36.

Mehrabian, A., and Russell, J. A. (1974). *An Approach to Environmental Psychology*. Cambridge, MA: M.I.T. Press, 266.

Midjourney (2023). *Midjourney (V5) [Text-to-image model].* Available online at: https://www.midjourney.com/

Open, AI. (2023). *DALL?E 2 [Text-to-image model].* Available online at: https://openai.com/dall-e-2

Pallasmaa, J. (2012). *The Eyes of the Skin : Architecture and the Senses.* Hoboken, NJ: John Wiley & Sons, 181.

Picard, R. W. (2003). Affective computing: challenges. *Int. J. Hum.-Comput. Stud.* 59, 55–64. doi: 10.1016/S1071-5819(03)00052-1

Pinaya, W. H., Tudosiu, P. D., Dafflon, J., Costa, P. F. D., Fernandez, V., Nachev, P., et al. (2022). "Brain imaging generation with latent diffusion models," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13609 LNCS, 117–126. doi: 10.1007/978-3-031-18576-2_12

Ploennigs, J., and Berger, M. (2023). AI art in architecture. *AI Civil Eng.* 2:8. doi: 10.1007/s43503-023-00018-y

Pressman, A. (2001). *Architectural Design Portable Handbook: A Guide to Excellent Practices*. New York, NY: McGRAW-HILL.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). "Learning transferable visual models from natural language supervision," in *International Conference On Machine Learning (PMLR)*, 8748–8763.

Schmidt, S. R., and Bohannon, J. N. (1988). In defense of the flashbulb-memory hypothesis: a comment on mccloskey, wible, and cohen (1988). *J. Exp. Psychol. Gen.* 117, 332–335. doi: 10.1037/0096-3445.117.3.332

Stability, A. I. (2023). DeepFloyd IF [Text-to-Image Model]. Available online at: https://www.deepfloyd.ai/deepfloyd-if

Sun, Y., Yang, C.-H., Lyu, Y., and Lin, R. (2022). From pigments to pixels: a comparison of human and AI painting. *Appl. Sci.* 12:3724. doi: 10.3390/app12083724

Tang, L., Yuan, P., and Zhang, D. (2023). Emotional experience during human-computer interaction: a survey. *Int. J. Hum. Computer Interact.* 1–11. doi: 10.1080/10447318.2023.2259710

Tao, J., and Tan, T. (2005). "Affective computing: a review," in *International Conference on Affective Computing and Intelligent Interaction* (Berlin; Heidelberg: Springer Berlin Heidelberg), 981–995. doi: 10.1007/11573548_125

Viliunas, G., and Grazuleviciute-Vileniske, I. (2022). Shape-finding in biophilic architecture: application of AI-based tool. *Arch. Urban Planning* 18, 68–75. doi: 10.2478/aup-2022-0007

Yildirim, E. (2022). "Text-to-image generation AI in architecture," in *Art and Architecture: Theory, Practice and Experience*, ed. H. Hale Kozlu (Lyon: Livre de Lyon), 97.

Zhang, Z., Fort Mir, J. M., and Mateu, L. G. (2022). The effects of white versus coloured light in waiting rooms on people's emotions. *Buildings* 12:1356. doi: 10.3390/buildings12091356

Zhang, Z., Fort, J. M., and Mateu, L. G. (2023). Exploring the potential of artificial intelligence as a tool for architectural design: a perception study using Gaudí's works. *Buildings* 13:1863. doi: 10.3390/buildings13071863

Zhao, S. (2016). "Image emotion computing," in *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference* (New York, NY: ACM), 1435–1439. doi: 10.1145/2964284.2971473

TYPE Original Research
PUBLISHED 20 March 2024
DOI 10.3389/fpsyg.2024.1300996

Implementing machine learning techniques for continuous emotion prediction from uniformly segmented voice recordings

Hannes Diemerling[1,2,3,4]*, Leonie Stresemann[4], Tina Braun[4,5] and Timo von Oertzen[1,2]

[1] Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany, [2] Thomas Bayes Institute, Berlin, Germany, [3] Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany, [4] Department of Psychology, University of the Bundeswehr München, Neubiberg, Germany, [5] Department of Psychology, Charlotte-Fresenius University, Wiesbaden, Germany

OPEN ACCESS

EDITED BY
Simone Belli,
Complutense University of Madrid, Spain

REVIEWED BY
Bagus Tris Atmaja,
National Institute of Advanced Industrial Science and Technology (AIST), Japan
Rajeev Ratna Vallabhuni,
Bayview Asset Management, LLC,
United States
Juan Manuel Mayor Torres,
University College Dublin, Ireland

*CORRESPONDENCE
Hannes Diemerling
✉ diemerling@mpib-berlin.mpg.de

RECEIVED 23 September 2023
ACCEPTED 09 February 2024
PUBLISHED 20 March 2024

CITATION
Diemerling H, Stresemann L, Braun T and von Oertzen T (2024) Implementing machine learning techniques for continuous emotion prediction from uniformly segmented voice recordings. *Front. Psychol.* 15:1300996.
doi: 10.3389/fpsyg.2024.1300996

Introduction: Emotional recognition from audio recordings is a rapidly advancing field, with significant implications for artificial intelligence and human-computer interaction. This study introduces a novel method for detecting emotions from short, 1.5 s audio samples, aiming to improve accuracy and efficiency in emotion recognition technologies.

Methods: We utilized 1,510 unique audio samples from two databases in German and English to train our models. We extracted various features for emotion prediction, employing Deep Neural Networks (DNN) for general feature analysis, Convolutional Neural Networks (CNN) for spectrogram analysis, and a hybrid model combining both approaches (C-DNN). The study addressed challenges associated with dataset heterogeneity, language differences, and the complexities of audio sample trimming.

Results: Our models demonstrated accuracy significantly surpassing random guessing, aligning closely with human evaluative benchmarks. This indicates the effectiveness of our approach in recognizing emotional states from brief audio clips.

Discussion: Despite the challenges of integrating diverse datasets and managing short audio samples, our findings suggest considerable potential for this methodology in real-time emotion detection from continuous speech. This could contribute to improving the emotional intelligence of AI and its applications in various areas.

KEYWORDS

machine learning (ML), emotion classification, audio emotion recognition, neural networks, speech signal features, Bilingual emotional classification

Introduction

Non-verbal communication, including the different aspects of a speaker's voice, plays a crucial role in conveying emotions and is highly valued in interpersonal interactions. While verbal content is important, research suggests that humans are significantly influenced by non-verbal cues, even in purely acoustic expressions of emotion (Miller, 1981). In an increasingly globalized world, where technical means of signal transmission have become essential, understanding emotions through non-verbal cues gains even more significance (Morton and Trehub, 2001).

Research suggests that one intriguing question arising in this context is whether technical tools are capable of accurately predicting mood or emotions based on vocal parameters and acoustic measurements, independent of semantic content. If so, then this could allow for the analysis of convergences and divergences between verbal and non-verbal expressions, enriching communication in various contexts.

Previous scientific research used semantically closed audio recordings of roughly 1.5–5 s to develop classification tools (Chen et al., 2018; Jiang et al., 2019; Mustaqeem and Kwon, 2019, 2021; Mustaqeem et al., 2020). However, to apply such tools to dynamically measure change in emotions, algorithms to analyze audio recordings that are not semantically restricted are needed. The objective of this article is to develop such a classification tool that can recognize emotions in the voice. The tool is designed to process audio recordings in 1.5 s segments, identifying emotions regardless of the semantic content of the audio.

The decision to process audio recordings in 1.5 s segments merits further explanation. Implementing fixed time windows serves a dual purpose. Firstly, it simulates real-life scenarios where audio clips may be randomly segmented without any predefined understanding of when an emotion begins or ends. By establishing an algorithm that classifies emotions from these fixed segments, we are ensuring that the tool is robust enough to process audio in various real-world applications. Secondly, the use of shorter, fixed windows is strategically designed to minimize the likelihood of capturing multiple or mixed emotions within a single segment. This will attempt to ensure that the emotional content of each clip is as pure as possible when using real data in the future, which should lead to a more accurate classification.

Our rationale for selecting a 1.5 s window specifically has both empirical and practical origins. Empirically, the work of Lima et al. (2013) provided insights into the feasibility of emotion recognition from short non-verbal vocalizations. In their study, participants exhibited high accuracy in predicting emotions from audio clips that averaged around a second in length, suggesting that meaningful emotional content can be discerned from relatively brief snippets of sound. Practically, the choice of a 1.5 s window is consistent with the nature of our dataset. The dataset, composed of audios ranging from 1.5 to 5 s, contains emotionally charged but semantically neutral sentences. By opting for a 1.5 s segmentation, we can ensure that nearly every audio segment retains its original length without the need to artificially lengthen it with added silence. This approach essentially aims to extract the most emotionally salient part of each recording, which in part corresponds to the short vocalizations described by Lima et al. (2013).

This article will evaluate different machine learning techniques for the development of a robust tool capable of classifying emotions using these 1.5 s long audio clips. The effectiveness of this tool will be compared with the human ability to recognize emotions through voice. If the accuracy of the developed classifier is comparable to human judgment, it could not only serve practical applications but also allow researchers to infer aspects of human emotion recognition through reverse engineering.

Decoding emotions

Contemporary emotion theories acknowledge the multidimensional nature of emotions, emphasizing their social and contextual aspects (Scherer, 2005; Fontaine et al., 2007; Moors et al., 2013). The tool presented in this article is based on Ekmans theory of basic emotions (Ekman, 1999). While Ekmans theory offers a practical and widely recognized framework, it is sometimes also criticized for its simplicity in representing human emotions. However, it provides a useful foundation for classifying emotions, while still allowing for a more nuanced understanding of emotions in future research.

TABLE 1 Classifier performance of studies using Emo-DB and RAVDESS databases.

Referenes	Method[a]	DB[b]	Perf.
Xiao et al. (2010)	NN-PC	E	81.2%
Chen et al. (2018)	CNN	E	82.8%
Jiang et al. (2019)	CNN	E	84.5%
Mustaqeem et al. (2020)	CNN	E	85.5%
Mustaqeem et al. (2020)	CNN	R	77%
Mustaqeem and Kwon (2019)	CNN	R	79%
Mustaqeem and Kwon (2021)	2S-CNN	E	95%
Mustaqeem and Kwon (2021)	2S-CNN	R	85%

This table provides an overview of accuracies achieved by various studies.
[a]Method: NN, Neural Network with pre-classification (PC); CNN, Convolutional Neural Network; 2S-CNN, two-Stream Convolutional Neural Network. [b]DB: E, Emo-DB; R, RAVDESS.

Emotions, as dynamic processes, encompass several interrelated components. The diverse manifestations of emotions at various levels can be classified based on their distinct patterns of expression. This article uses the definition by Goschke and Dreisbach (2020) which includes all relevant parameters, giving a holistic picture of the multifaceted nature of emotions:

> "Emotions are psychophysical reaction patterns based on more or less complex evaluations of a stimulus situation, which are accompanied by a series of peripheral physiological changes as well as the activation of certain central nervous systems. These reactions motivate certain classes of behavior, can be expressed in specific facial expressions and body postures, and are often (but not necessarily) associated with a subjective quality of experience" (Goschke and Dreisbach, 2020).

This article follows the assumption that emotions, despite their nature as dynamic processes consisting of multiple components, can be assigned to categorize based on their patterns of expression. This assumption follows the concept of basic emotions, which Scherer (1985) recognizes as the main types of emotions. Ekman (1999) specifies the seven basic emotions as fear, surprise, anger, disgust, joy, sadness, and contempt, which have universal characteristics and are intuitively performed and also recognized by humans.

The ability to recognize and classify emotions is called cognitive empathy. Not every emotion is recognized equally well, as cognitive empathy is a combination of many subskills with interpersonal and intrapersonal differences (Marsh et al., 2007). In a conversation, not only linguistic cues are used to recognize emotions, but also non-verbal paralinguistic cues. Paralinguistic signals accompany what is spoken, for example, speaking rate or volume, and expands the spoken words with additional aspects that provide information about the speaker's state of mind (Bussmann and Gerstner-Link, 2002).

Emotions in the voice

The facial and vocal expression of basic emotions are understood cross-culturally, and these emotions are associated with similar physiological patterns of change (Ekman et al., 1983). These emotions are also universally recognized through vocal expression (Izdebski, 2008). The human voice serves as a powerful channel for expressing emotional states, as it provides universally understandable cues about the sender's situation and can transmit them over long distances. Voice expression is rooted in brain regions that evolved early in human development, underscoring its fundamental role in our evolutionary history (Davitz, 1964; Morton, 1977; Jürgens, 1979).

When categorizing emotions based on vocal expressions, employing a limited number of emotion categories proves advantageous to avoid overwhelming information (Johnson-Laird and Oatley, 1998). Additionally, distinct emotion specific patterns of acoustic features have been observed (Scherer, 1979), which can still be detected even after removing linguistic cues from the speech signals. Physiological parameters significantly influence vocal parameters like loudness, fundamental frequency, noise components, and timbre (Trojan et al., 1975; Frick, 1985; Burkhardt, 2000).

Related publications

Several classification tools have been developed to recognize and classify emotions in the voice. A notable example is Xiaos classifier, which utilizes artificial neural networks and incorporates pre-classification to enhance accuracy (Xiao et al., 2010). More recent developments have focused on convolutional neural networks (CNNs) and their ability to efficiently process large amounts of data (Chen et al., 2018; Jiang et al., 2019; Mustaqeem and Kwon, 2019, 2021; Mustaqeem et al., 2020). For instance, the study by Mustaqeem and Kwon (2021) introduces a complex two-stream CNN that achieves high accuracies for different emotion databases. Table 1 presents the performance metrics of various classification tools as reported in the cited studies, which utilize differing methodologies:

1. Xiao et al. (2010) employ a 10-fold cross-validation method with a 50:50 train-test split for each fold and include a preclassification step to determine gender. The table lists their reported average accuracy.

2. Chen et al. (2018) implement a 10-fold cross-validation, splitting the data for each split by speakers: eight for training, one for testing, and one for validation, targeting four emotional states happy, angry, sad, and neutral. The corresponding average accuracy figures are depicted in the table.

3. Jiang et al. (2019) adopt a Leave-One-Speaker-Out (LOSO) approach. Shown in the table is the unweighted average accuracy accumulated across all trials.

4. Mustaqeem et al. (2020) use a 5-fold cross-validation, designating eight speakers for training and two for testing in each fold. The table illustrates their average accuracy results.

5. Mustaqeem and Kwon (2019) execute a 5-fold cross-validation with an 80:20 split for training and testing, respectively. Their average accuracy is shown in the table.

6. Mustaqeem and Kwon (2021) perform a 10-fold cross-validation with an 80:20 train-test split, with the table showing the F1 scores as the most relevant performance parameter, as presented in the referenced source.

However, it is important to note that the performances outlined above cannot be directly compared with the results of this article. Firstly, the methodologies employed across these studies vary. Secondly, the databases used are also distinct, given that this study utilizes audio clips trimmed to 1.5 s as opposed to complete audio recordings. In particular, we aim to demonstrate that emotion recognition based on voices, when using the right tools, is also possible when using very short time segments, which can be used for continuous emotion classification of voice data. The performances shown are intended to provide an overview of the existing classifiers that have been trained on the data used here in order to be able to better contextualize this article.

All the aforementioned approaches utilize audio recordings from the Emo-DB and the RAVDESS databases. These databases offer clearly recognizable emotion recordings in complete sentences or uniformly defined speech units, which has led to limited attention being given to audio segmentation in previous research. However, the challenge lies in spontaneous speech where defining unambiguous units becomes difficult. An effective segmentation approach needs segments long enough to extract acoustic patterns but also short enough to capture emotional state changes. Studies on continuous segmentation have already been undertaken in the literature. Atmaja and Akagi (2020) showed emotion recognition beyond chance for a visual-auditory dataset using a 4 s time window.

Contrary to the studies mentioned above, the work of Stresemann (2021) takes a different approach. She standardized all audio recordings from these databases to a length of 1.5 s, analyzing them as independent units without considering the grammatical sentence structure. The aim is to focus purely on emotion recognition, disconnecting it from the semantic content of the sentences. This choice of approach, which sometimes results in the cropping of longer files and the potential loss of words, is supported by Scherer (1979). He argued for the existence of emotion-specific acoustic patterns that are independent of contiguous sequences. This approach not only aids in mapping emotion expression changes within longer sentences but also has a practical benefit: it is especially applicable in online settings where smaller datasets can be quickly analyzed, and reliable assessments can be made.

Approach of this study

This article aims to enable automatic continuous classification by limiting the duration of individual audio segments to 1.5 s. The practical objective is to continuously split a longer audio track into potentially overlapping sequences, allowing the model to provide a continuous assessment of emotions in the voice. The study by Stresemann (2021) serves as the foundation for this article, but the approach here uses a more automated method with advanced machine learning techniques. The fixed time length of 1.5 s is intended to simulate the challenges in real-life datasets. Using audio files of different lengths would require upstream recognition in real data. The specific length of 1.5 s serves as a compromise between the shortest possible audio length to avoid overlapping emotions and enough information to still allow humans to understand the audio files. The aim of this study is to proof that automatic classification of human speech is possible under these constraints. Thereby, we aim to show that a tool can be created, which automatically classifies emotions in continuous human speech, without the necessity of elaborate preprocessing. To do so, we present an approach that compares different model designs and different combinations of linguistically diverse audio tracks in terms of their accuracy in emotion recognition both to each other and to humans.

Methodology

Building on the theoretical background outlined earlier, this section delves into the methodology employed in this study. The processing of the audio data is discussed first, followed by a detailed explanation of the datasets and a comparison with human performance. The latter part of this section will describe the generation of individual features and the development and testing of various models.

Audio

The audio material for this study was sourced from two publicly accessible emotion databases from distinct cultures: Germany and Canada. This choice is grounded in the cross-cultural universality of emotions in audio, as supported by the meta-analysis conducted by Juslin and Laukka (2003). The considered emotions for this study include joy, anger, sadness, fear, disgust, and neutral.

Specifically, English-language recordings were extracted from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS; Livingstone and Russo, 2018). An example of content from RAVDESS is the neutral statement, “Dogs are sitting by the door.” For German-language recordings, the Berlin Database of Emotional Speech (Emo-DB) was used (Burkhardt et al., 2005). A representative sentence from Emo-DB is “Der Lappen liegt auf dem Eisschrank” (the rag lies on the refrigerator). In both databases, actors induced the emotions using emotional memory techniques.

For the audio processing stage, we settled on a strategic duration of 1.5 s per segment. This choice was influenced by several factors: to emulate real-world conditions where snippets of emotion may lack clear starting or ending points, to approximate the briefest discernible emotional span, and to minimize the potential for overlapping emotions in a single clip. Files longer than this were trimmed to capture the core 1.5 s, with any excess equitably truncated from both the start and end. Conversely, shorter files were symmetrically extended with silence on both sides, ensuring a consistent segment length while preserving the original emotional content. In other studies (e.g., Chen et al., 2018; Jiang et al., 2019; Mustaqeem and Kwon, 2019; Mustaqeem et al., 2020), the audio files were not segmented. In order to additionally examine whether and how much accuracy is lost due to the selected length of the segmentation of the audio recordings, audio files that were segmented to 3 or 5 s were also used for parts of the utilized model designs. The same segmentation method was used for all variants.

The Ryerson Audio-Visual Database of Emotional Speech and Song

The RAVDESS is an open-access database offering 7,256 English-language recordings, both spoken and sung, spanning across three modalities: audiovisual, video-only, and audio-only (Livingstone and Russo, 2018). For the purpose of this study, only the audio modality was employed. Featuring recordings from 24 actors (12 male, 12 female), the database represents six emotions (joy, sadness, anger, fear, surprise, and disgust) in addition to two baseline states (neutral and calm). From RAVDESS, this research incorporated 1,056 audio clips, omitting the emotions of surprise and calm, each trimmed to a precise duration of 1.5 s.

Berlin Database of Emotional Speech

The Emo-DB, hosted by the Technical University of Berlin, is a public database comprising 535 German-language recordings, conducted by 10 actors (five male and five female) under the guidance of phoneticians (Burkhardt et al., 2005). The database encompasses the emotions of anger, fear, joy, sadness, disgust, and neutral speech. From the Emo-DB, 454 recordings were incorporated into this study, with the emotion of surprise excluded, and every clip was trimmed to 1.5 s.

Comparison to human performance

The data format for this research aligns with the methodology of Stresemann (2021), involving 61 participants (36 male and 25 female) aged between 20 and 71 years. Participants were tasked with a forced-choice format survey where they matched emotions to 82 English language recordings from the RAVDESS database and 88 German recordings from the Emo-DB. Covered emotions were fear, anger, joy, sadness, disgust, and neutral speech.

Before starting, participants received comprehensive information regarding the study procedure, data privacy guidelines, and the voluntary nature of participation. The survey also collected demographic details, including sex, age, first language, current domicile, and prior experience in English-speaking regions. The listening exercise required a quiet environment, where participants identified emotions immediately after a single playback. In cases of unclear recordings due to technical issues, an alternative “no statement” option was available. All data, barring one problematic disgust recording, were included in the final analysis.

The findings of Stresemann (2021) revealed a robust positive correlation between recognition rates on the Emo-DB and RAVDESS databases, indicating that individual empathic abilities might supersede linguistic or cultural biases in emotion recognition. This correlation is possibly influenced by the shared Germanic roots of English and German, leading to similarities in fluency and intonation. Conversely, studies contrasting different linguistic backgrounds highlighted advantages for listeners when the recordings matched their native tongue. For instance, native

English speakers surpassed Spanish and Japanese counterparts in emotion recognition (Graham et al., 2001). Similarly, Korean speakers outdid their French and American peers when classifying emotions in Korean (Chung, 2000).

While basic emotions' expression is broadly universal, nuances exist due to cultural differences (Graham et al., 2001). However, numerous studies, such as Juslin and Laukka (2003), underscore high cross-cultural emotion recognition rates. This suggests that even amidst cultural distinctions in emotional expression, humans' inherent auditory-driven emotion recognition abilities transcend linguistic and cultural confines. This inherent capability, albeit less refined than facial emotion recognition, does not necessitate formal training or guidance.

Feature generation

Once audio recordings were streamlined into 1.5 s segments, we embarked on generating a diverse set of features. The ambition was to mine maximum information through various methodologies, ensuring redundancy was at its minimal.

The following is an overview of the individual features created in this study. Each feature was calculated for each audio recording. For some features, summary values for the 1.5 s were computed (e.g., the mean for pitch). Table 2 gives an overview of all features together with the number of data points this feature provides. The "Summarization" column describes the summary approach for each variable, if one was used. Overall, there were 14,244 different entries for each audio recording. Given the potentially multidimensional nature of expressing each emotion in the voice, preselecting features could result in information loss. Therefore, the approach in this study was to generate as many features as possible, allowing the models to independently select relevant features. The features used here include:

Unmodified Audio Signal

The Unmodified Audio Signals served as the foundation for all subsequent feature calculations. A portion of the signal was preserved to retain potential unbiased information that may not be captured by other features.

Spectral flatness

Spectral Flatness is a measure of how evenly the energy of an audio signal is distributed across different frequency bands compared to a reference signal. It provides an estimate of the flatness of the signal and may be associated with certain emotions (Dubnov, 2004).

Spectral centroid

The Spectral Centroid indicates the average frequency at which the energy of a sound signal is centered. It can be used to estimate the perceptual brightness or tonal brightness of the sound and is sometimes related to valence and arousal, which are closely connected to emotions (Klapuri and Davy, 2007).

Fundamental frequency

Fundamental Frequency (F0) estimation means determining the lowest frequency and rate of periodicity in a sound signal. Analyzing the F0 provides information about the emotional dimensions of the signal (Cheveigna and Kawahara, 2002; Mauch and Dixon, 2014).

TABLE 2 Enumeration of dataset features, summarization, and quantity.

Feature	Summarization	Quantity
Unmodified Audio Signal	Variance	1,200
HPSS	Variance	2,400
Spectral Flatness	N/A	47
Spectral Centroid	N/A	47
Fundamental Frequency	N/A	47
Spectral Rolloff	N/A	94
Spectral Bandwidth	N/A	47
Zero Crossing Rate	N/A	47
Root Mean Square	N/A	47
Spectral Contrast	N/A	188
Tonnetz	N/A	282
Chroma	N/A	564
Pitch Tracking	Var. and mean*	2,050
Pitch Magnitudes	Var. and mean*	2,050
Magnitude	Var. and mean*	2,050
Phase	Var. and mean*	2,050
MFCC	N/A	940

Features, summarization, and quantity for the dataset. *Variance and mean calculated for each 2,048 Hz window.

Voiced

In addition to F0 estimation, the presence of a voice within a specified time window of the audio was measured, along with the probability of voice presence. The specific time window used was 2,048 Hz (Cheveigna and Kawahara, 2002; Mauch and Dixon, 2014).

Spectral rolloff

Spectral Rolloff indicates the frequency level at which a certain percentage (here, 0.85) of the energy is contained in the signal. It can identify the frequency ranges that are most strongly represented in the signal and may aid in emotion recognition (Sandhya et al., 2020).

Pitch tracking

Pitch Tracking estimates the pitch or fundamental frequency (F0) of a sound signal and measures its magnitude. This feature can provide additional information related to the F0 and assist in emotion classification (Smith, 2011).

Harmonic percussive source separation

The HPSS technique separates a sound signal into its harmonic and percussive components. Both components could convey different emotional information (Fitzgerald, 2010; Driedger and Müller, 2014).

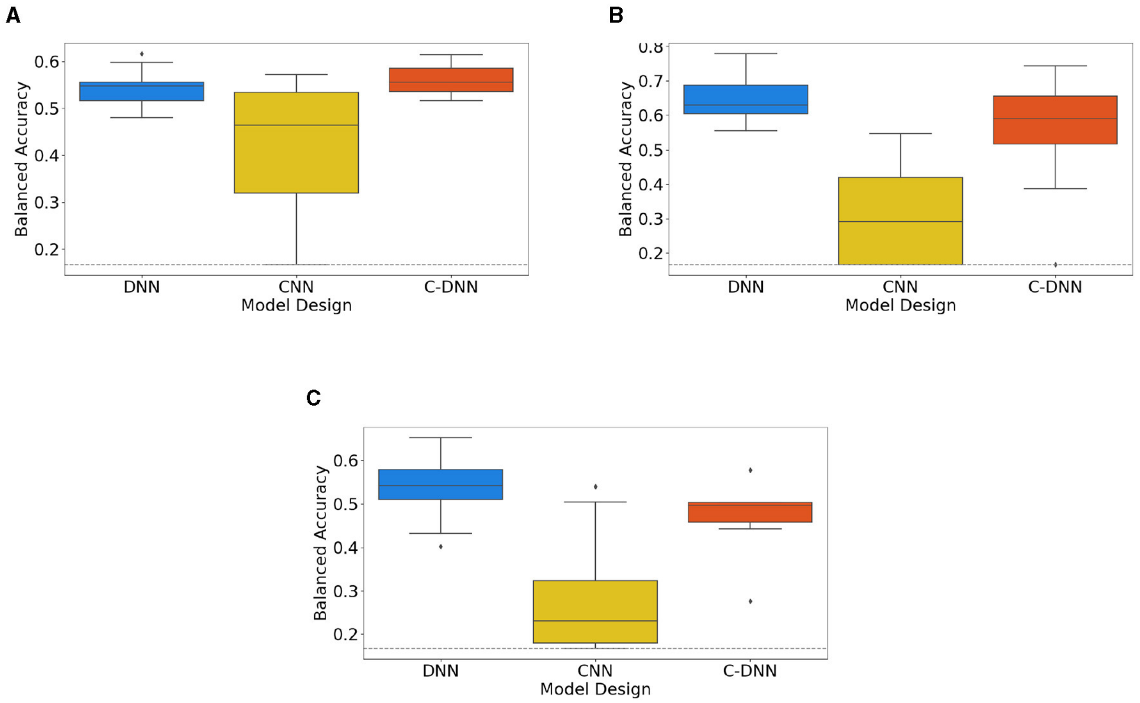


FIGURE 1
Neural network design (DNN, CNN, and C-DNN) comparisons based on cross-validation results. This figure shows the results of 10-fold cross-validations for the comparison between different neural network designs (DNN, CNN, and C-DNN) based on both combined and separate datasets. Subfigures represent: **(A)** results based on the combined Dataset, **(B)** results based on Emo-DB, and **(C)** results based on RAVDESS. The gray dashed line indicates the Balanced Accuracy of a random classifier.

Magphase

Magphase separates the complex-valued spectrogram (D) into its magnitude (S) and phase (P) components, where D = S × P. The magnitude is used to calculate various emotion-related features presented in this section, while the phase angle is measured in radians and used as is. The phase encodes relationships between different frequency components of the signal, which may contain emotional information, although it is rarely used in emotion classification (Librosa Development Team, 2023).

Spectral bandwith

Spectral Bandwidth is a measure of the spread of the spectral content of the signal. It is related to the frequency range of the signal and may be relevant to emotions (Klapuri and Davy, 2007).

Spectral contrast

Spectral Contrast refers to the differences in energy levels between different frequency ranges of an audio signals. It can describe the tone color of a signal, which might be associated with certain emotions (Jiang et al., 2002).

Zero crossing rate

The Zero Crossing Rate indicates the number of times the signal changes from positive to negative or vice versa. It can provide information about the dynamics of the signal (Hung, 2009).

Mel-frequency cepstral coefficients

MFCC are widely used features in music and speech recognition. They represent the Mel-requency energy distribution of an audio signal and can identify the most important frequencies of the signal while being robust to changes in loudness and sound characteristics (Sato and Obuchi, 2007).

TABLE 3 The mean of balanced accuracies of various models based on 10-fold cross-validation.

Dataset	DNN	CNN	C-DNN
Combined	54.49%	41.56%	56.24%
Emo-DB	64.69%	30.68%	54.85
RAVDESS	53.55%	28.39%	48.09%

DNN, Deep Neural Network; CNN, Convolutional Neural Network; C-DNN, Combination of Deep Neural Network and Convolutional Neural Network.

Root mean square

RMS is a measure of the average power of an audio signal. It indicates the average loudness of the signal and can describe its loudness level (Chourasia et al., 2021).

Tonnetz

Tonnetz is another representation of frequency ranges that can be used to identify the harmony of a musical signal, which might be associated with certain emotions (Harte et al., 2006).

Chroma

Chroma represents the presence of different frequency ranges in a music signal and can be used to identify the key of the music signal, potentially containing emotion-related information (Ellis, 2007).

Creation of the spectrograms

Spectrograms visually depict the frequency spectrum of audio signals, reflecting energy distribution across time and frequency. Such patterns have been identified as crucial in emotion recognition (Kim et al., 2010). For our study, spectrograms were crafted for every audio recording, saved as PNGs (without axes or borders) at a resolution of 320 × 240 pixels.

Subsequently, we detail the employed classification models.

The deep neural network

DNNs, renowned for their prowess in intricate pattern recognition, consist of interconnected feedforward layers with varying neuron counts (LeCun et al., 1998). The architecture allows the model to adjust to input data, predicting emotions via gradient-based learning.

The convolutional neural network

The generated spectrograms consist of numerous data points, resulting in 230,400 data points (320 × 240 × 3) for each image. To efficiently analyze these images, CNNs are employed. These networks, skilled at image processing through local receptive fields and weight sharing, enhance the representation using pooling, particularly max pooling, to retain essential data while reducing the image size (LeCun and Bengio, 1995).

The hybrid model C-DNN

Our hybrid C-DNN model merges the insights of both the generated features and spectrograms. It encompasses a dual-input approach: a DNN for feature processing and a CNN for spectrogram analysis. The output layers from both networks converge into a concatenated layer, followed by another feedforward DNN predicting emotions through a softmax function. The goal is to determine whether combining spectrograms and features improves information extraction compared to individual data sources.

Creation of the models

The three model designs described above were implemented in a Python environment using Tensorflow (Abadi et al., 2015) and Scikit-learn (Pedregosa et al., 2011). The dataset was apportioned into training (80%) and test sets. The hyperparameters for each model were defined separately using Bayesian optimization with a Gaussian process based on the associated training dataset. A brief overview of the hyperparameter is listed in Table A1. Using Bayesian optimization, different models were formed, their hyperparameters adjusted, and subsequently trained on the training dataset. Post every training epoch, the test dataset underwent a prediction process. After completing up to four training epochs, validation accuracy was gauged a final time. The validation accuracy from the test data was then used as a benchmark to avoid overfitting.

Testing the different models

For a more consistent comparison with existing literature, the models underwent a 10-fold cross-validation.

The performance metrics employed to measure the model quality included Balanced Accuracy (BAC). This was compared to both random classifications and the BAC achieved by other models.

Our evaluation approach combined Independent Validation (Kim and von Oertzen, 2018) with Bayesian Updating. Initially, models were trained on 10% of the total data, setting aside another 10% for validation, ensuring overfitting was kept within limits. The models were then sequentially introduced to new data in chunks of 16 data points. Before integrating these data points into the primary training dataset, the models attempted their prediction, updating the BAC's posterior distribution via Bayesian techniques. This cyclic procedure continued until the entire dataset had been incorporated into the training set, with the validation set consistently monitoring for overfitting.

Successful and unsuccessful predictions were used to update the parameters of a beta distribution through Bayesian Updating, providing a posterior distribution of the classifier accuracy. A beta distribution was chosen to model the accuracies as it can depict that a perfect accuracy of 1 is very unlikely or even impossible, while other values can be equally likely. By comparing the overlap between the beta distributions of the models, one could assess the probability of one model outperforming another, for instance, a classifier that merely guesses the results. This statistical approach allowed us to validate the effectiveness and generalizability of our model while providing a measure of uncertainty.

Testing against humans

To evaluate the performance of human participants, we used a similar approach, assuming a binomial distribution for the correct recognition of emotions. We then estimated the accuracy using a beta distribution. By comparing the overlaps among the distributions for each emotion, we can determine the similarity in performance and assess the likelihood of differences between human participants and the classifiers.

Results

This section presents the outcomes from the model comparisons. First, we compared the models using cross-validation. For all following Bayesian accuracy estimations, we used a beta(1,1) prior, which stands as the conjugate prior for a binomial distribution, representing minimal prior information.

Cross validation

Figure 1 presents the outcomes of 10-fold cross validations for three distinct model designs, individually trained on different datasets: the combined dataset in A, the Emo DB dataset in B, and the RAVDESS dataset in C, respectively. The boxplots illustrate the model performances, offering a visual comparison across the diverse model designs and datasets. The

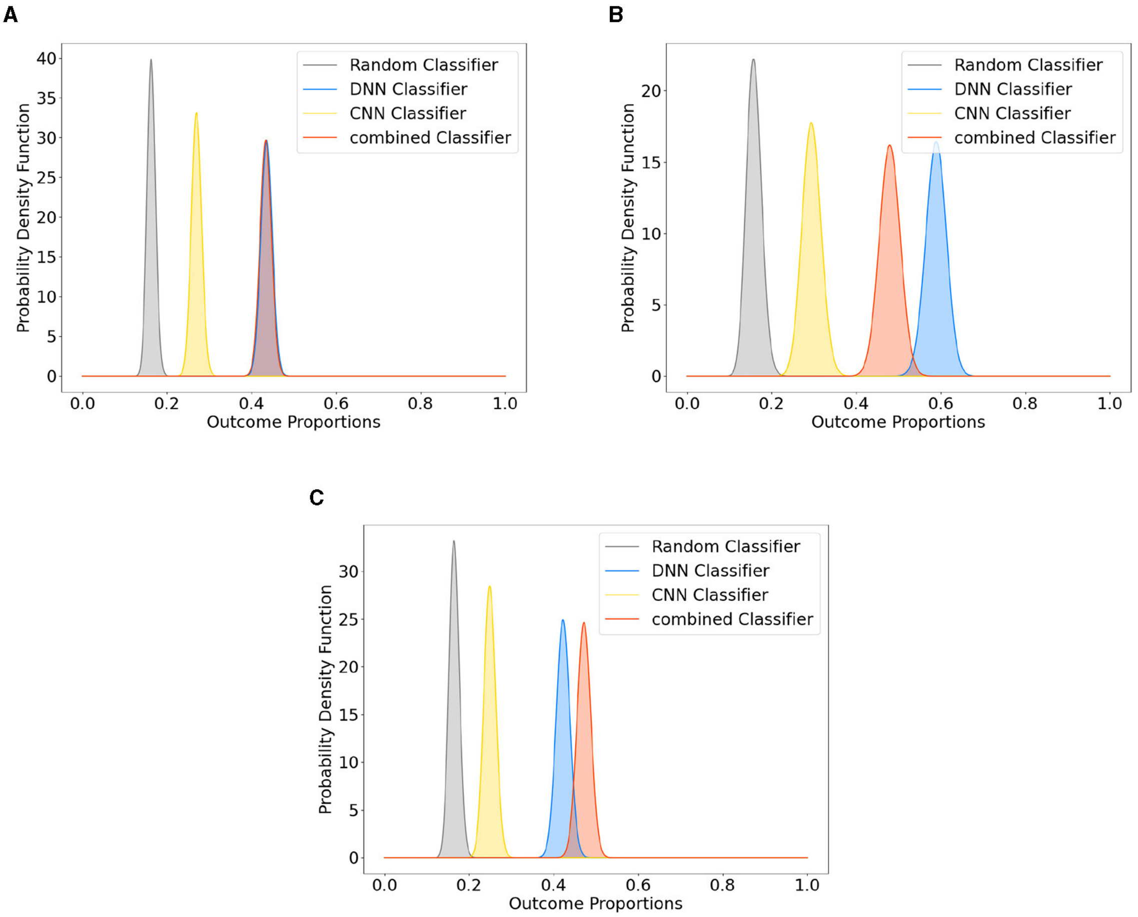


FIGURE 2
Posterior distributions of neural network designs (DNN, CNN, and C-DNN) vs. a random classifier. This figure demonstrates the posterior distributions for different neural network designs (DNN, CNN, and C-DNN) compared to a random classifier. Subfigures represent: **(A)** Classifier comparison based on the Combined Dataset, **(B)** Classifier comparison based on Emo DB, and **(C)** Classifier comparison based on RAVDESS.

corresponding mean values for each model design, computed from the different datasets, are consolidated in Table 3, thereby facilitating a numerical evaluation of the model performances. For the model design DNN, additional models are created based on 3 and 5 s segmented audio files. This results for the combined dataset in 62.36% (3 s) and 61.79% (5 s). For the Emo-DB dataset, the results are 72.91% (3 s) and 69.21% (5 s). Results for the RAVDESS dataset are 60.01% (3 s) and 61.00% (5 s).

Combined dataset

Figure 2 presents the results obtained from the Bayesian estimate of the classifier accuracies. The three different model designs, that is, DNN, CNN, and combined, all trained on the Emo DB and RAVDESS datasets combined. The posterior distribution of each classifier is shown alongside the posterior distribution of random classification. The posterior distributions indicate where the true performance under each classifier is expected to be. A distribution closer to the maximum value of 1 indicates a better performance. The posterior distribution of the random classifier (indicating guessing) is to the left of the posterior distributions of the trained classifiers and only overlapps by 1%. This indicates that the probabilbiity that the classifiers perform better than guessing is above 99%. The position of the distributions is described by the maximum a-posteriori estimate (MAP), the peak of the posterior distribution. The MAP performance of two of the models (DNN and C-DNN) is close to 0.45 (0.436 and 0.433) with a standard error of 0.013. The CNN model performance is lower compared to the other models (0.27) with a standard error of 0.012. Note that with six categories to classify, guessing performance is 1/6. Analysis of the average saliency maps across all spectrograms obtained from the Emo DB, RAVDESS, and combined datasets has provided insights into the time-segment relevance for emotion classification. As depicted in Figure 3, the distribution of SHAP values across 48 time segments reveals variations in the predictive importance of certain time intervals. Notably, segments with higher SHAP values indicate a stronger influence on model predictions, which suggests that certain temporal portions of the audio recordings are more salient for emotion detection.

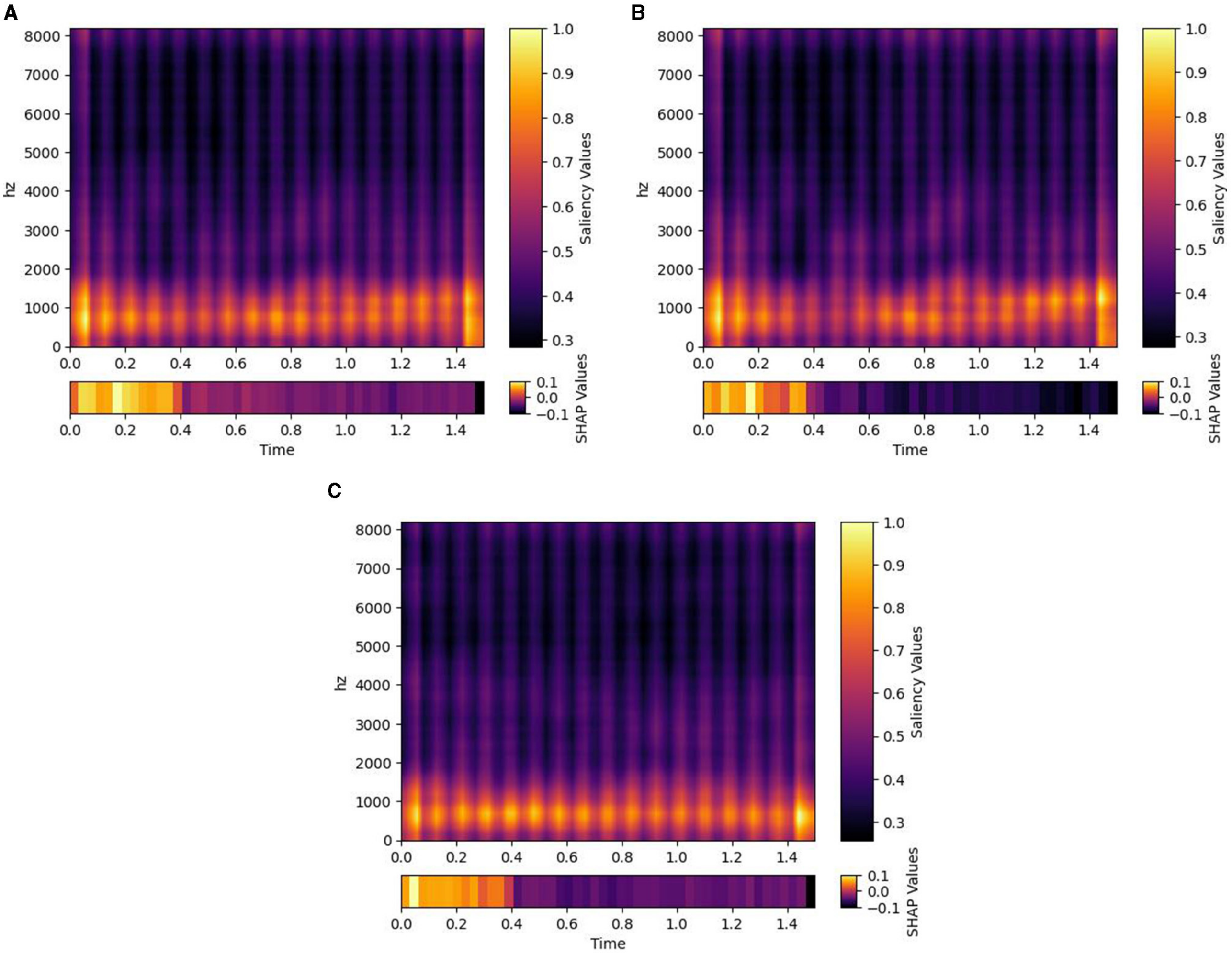


FIGURE 3
Saliency maps and SHAP values for different datasets. Each plot illustrates the average saliency across all spectrograms derived from emotional audio recordings within their respective datasets. **(A)** represents the combined dataset, **(B)** the Emo_DB dataset, and **(C)** the RAVDESS datasets. The color gradient within each plot signifies varying saliency values, while the bar beneath it provides SHAP values for 48 time segments, indicating the significance of individual features grouped by time intervals.

Emo DB separated

To further investigate the performance of the Emo DB and RAVDESS datasets separately, a corresponding method was used to compare them to a random classifier. The corresponding results for the Emo DB Dataset are shown in Figure 2B. The analysis shows that assuming a flat prior, the probability of the models on these datasets differing from a random classifier is over 99%. The posterior distribution of each classifier is shown alongside the posterior distribution of random classification. The MAP performance of two of the models (DNN and C-DNN) is close to 0.5 (0.58 and 0.48) with a standard error of 0.024. The CNN model performance is lower than the other models (0.29) with a standard error of 0.022.

RAVDESS separated

The corresponding results for the RAVDESS Dataset are shown in Figure 2C. The posterior distribution of each classifier is shown alongside the posterior distribution of the random classification. The probability that the classifier performs better than guessing is above 99% throughout. The MAP performance of all three models (DNN, CNN and C-DNN) is close to 0.5 (0.42 and 0.42) with a standard error of 0.016. The CNN model performance is lower than the other models (0.26) with a standard error of 0.014.

Comparison to humans

Figure 4 presents a comparative analysis between the three model designs and human performance in classifying the basic emotions and neutral. Each sub-figure corresponds to an emotion, namely, fear A, joy B, anger C, disgust D, sadness E, and neutral F. Both the DNN and the C-DNN design show comparable performance with the participants while the CNN shows unreliable performance across emotions. The sub-figures illustrate the beta distributions of the classifiers' performance. The spread and central

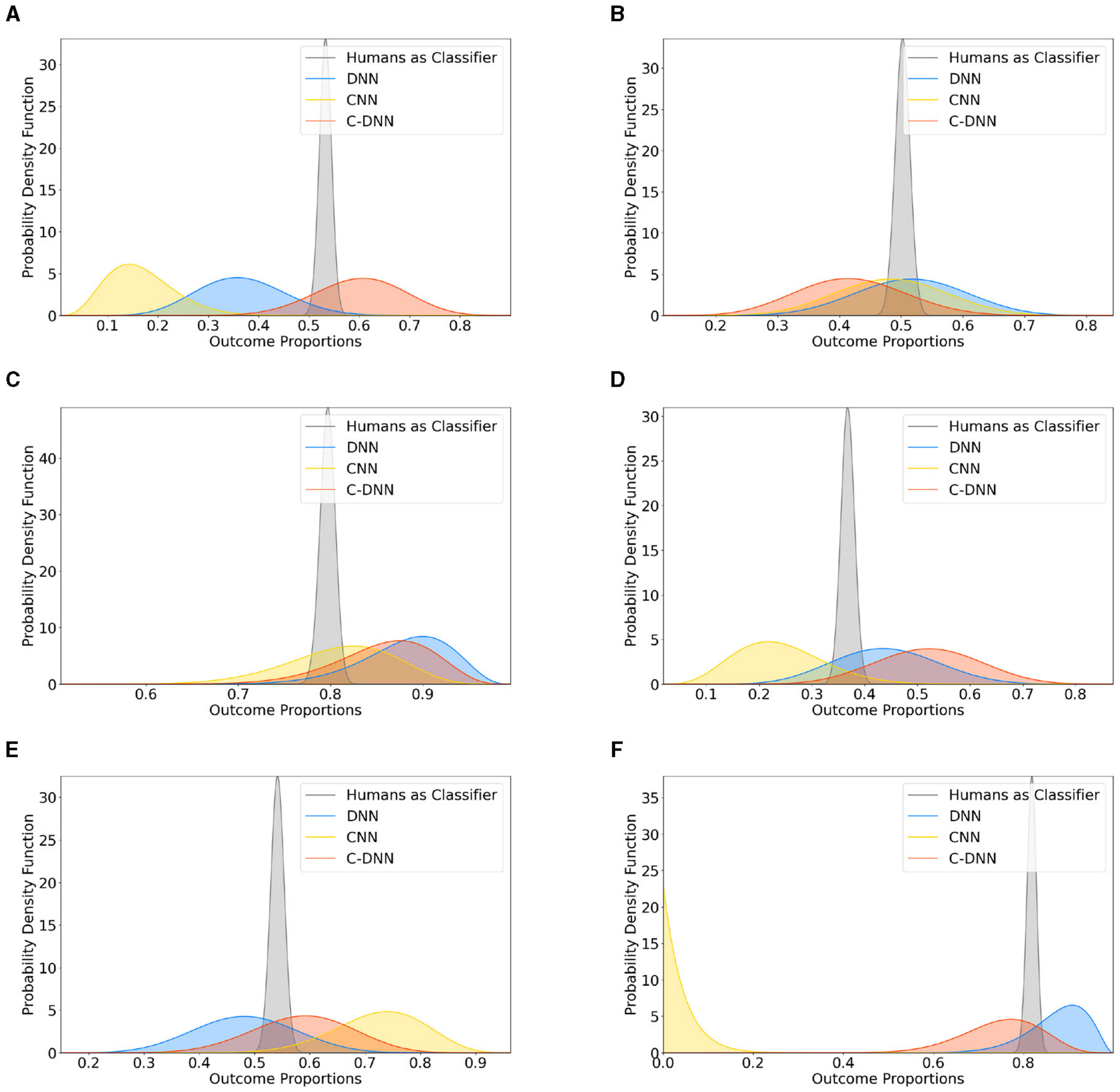


FIGURE 4
Different neural network designs (DNN, CNN, and C-DNN) compared to human classification across emotions. This figure presents the updated beta distributions for the comparison between different neural network designs (DNN, CNN, and C-DNN) and humans in classifying different emotions. Subfigures correspond to: **(A)** fear, **(B)** joy, **(C)** anger, **(D)** disgust, **(E)** sadness, and **(F)** neutral.

tendencies of these distributions provide an understanding of the variance in the performance of the models and the humans.

Discussion

This article compared the effectiveness of three model designs: Deep Neural Network (DNN), Convolutional Neural Network (CNN), and a combination of the two (C-DNN). Each model was trained and evaluated using three different versions of datasets. The methods of evaluation included 10-fold cross-validation, a combination of Independent Validation and Bayesian Updating, and a comparison with human performance.

The cross-validation revealed the combined model (C-DNN) to be most effective on the combined dataset, while the CNN showed less performance and reliability across all datasets. When a combination of Independent Validation and Bayesian Updating was used, each model performed notably better than random guesses. Nonetheless, the CNN model showed lower performance than its counterparts under all circumstances.

A comparison with human emotional state classification revealed that the DNN and C-DNN models performed at a level similar to humans, whereas the CNN model was less consistent across all emotions.

Design-specific aspects

The CNN model design in this article showed strong overfitting, leading to poorer and less stable performance than anticipated. This overfitting could be attributed to the segmentation of audios into 1.5 s units, which may have disrupted the emotional structure and limited the models ability to capture nuanced emotional patterns. Future research should explore improving approaches that capture the temporal dynamics of emotions more effectively. For instance, using overlapping windows might be beneficial. This approach would involve half-second increments of audio, providing significant overlap to average out effects. This could potentially capture varying emotional patterns more effectively, even beyond the usual 1.5 s segments.

Different datasets

When comparing the different datasets, it is evident that all models can predict emotions based on the generated features from audios better than guessing and in the case of the DNN and C-DNN comparatively well as humans. However, the performances vary. The Emo DB dataset consistently leads to the best performance for the DNN design and also for the C-DNN design by excluding the one outlier. It is important to note that this dataset is smaller and less diverse compared to the RAVDESS dataset. Therefore, better generalization of the models cannot be derived from higher performance.

RAVDESS and Emo DB combined

The combined dataset from Emo DB and RAVDESS produced comparable performances to the results based on the RAVDESS dataset. Only the CNN design showed inconsistent results, while the other two models showed consistant results.

Although English and German share a common Germanic origin, a uniform language-specific emotional expression cannot be assumed. The consistent performance for the DNN and CNN designs is, therefore, even more remarkable. Despite the limitations of the clipped audio recordings and the heterogeneous datasets, they show a consistent performance across the different datasets used. In particular, considering the comparable performance to the participants, it could be argued that the models have recognized the underlying patterns of emotion contained in the audio recordings beneath the culturally specific facets.

RAVDESS and Emo DB separated

It is worth noting that the RAVDESS dataset is significantly larger than the Emo DB dataset and consists of English recordings with a neutral North American accent, which includes Canadian English. Canadian English is characterized by "Canadian Rising," a phenomenon that affects vowel formants and could impact the acoustic analysis and emotion recognition accuracy. It describes the habit of pronouncing vowels that are normally pronounced with low tongue position with a middle tongue position (Chambers, 1973). The key point is not the change in the word's pronunciation (where vowels sound higher) but the accompanying shift of the vowel formants. This linguistic phenomenon is visible in the acoustic analysis and could thus cause slight irritations with regard to emotion recognition, which are reflected in the performance of the classifier. This aspect could also be a limiting factor for the models based on the combined dataset, as it could prevent further generalization.

Different models

An integral part of this article was an investigation into whether the combined C-DNN model, leveraging both spectrograms and numerical features, could offer additional informational benefits over the DNN and CNN models used independently. The C-DNN model did exhibit a minor improvement in performance; however, this incremental gain did not proportionally reflect the potential combination of the DNN and CNN models, as one might intuitively expect. This suggests that the added complexity of the C-DNN may not necessarily translate into substantial gains in emotion recognition performance. One possible explanation is that the information in the spectrograms might already be represented in the generated features. Consequently, the additional data from the spectrograms might not enhance the generalization of emotion recognition. Also, the cropping of the audios could have reduced the information value of the spectograms to such an extent that they can no longer reliably represent emotion-related information. Both of these aspects could contribute to the CNN design over-adapting to non-emotion-related aspects or learning culture-specific facets lacking compensation from the features unlike the C-DNN.

Comparison with previous studies

A classification based on short 1.5 s audio sequences has not been approached in the literature to the authors best knowledge. Short clips of this length are a solution approach when it comes to classifying the emotions to be heard within a longer audio stream without performing complex preprocessing. As can be seen in Table 1, performance for longer audio sequences (in the literature listed there, ranging from 1.5 to 5 s) can allow for higher accuracies. We have deliberately worked with audio files as brief as 1.5 s to highlight the feasibility and potential of real-time emotion recognition in dynamic settings. Longer audio clips might yield more accurate results; however, they are less reflective of actual conditions where audio data is rarely perfect and manually segmenting emotional content is often unfeasible. Our choice of a 1.5 s timeframe aims to emulate an automated system that may imperfectly trim audio segments, thereby mirroring the practical challenges faced by classifiers in real-world applications. These segments are short and concise enough for human comprehension and also represent the minimal length necessary to retain substantial information from the raw audio without introducing uninformative content into the analysis. In addition, models were created for the DNN designs based on differently segmented audio files (3 and 5). As expected, there is a higher accuracy for the 3 s audio files, but no clear increase for the 5 s length. This could be due to the type of audio processing, as the audio files that were too short were lengthened by adding silence. This could, on the one hand, make the classification more difficult and, on the other hand, could require a higher complexity of the models in order to learn the correct patterns. This additional complexity could potentially require more computing power than was employed. It should be highlighted that in this article good performances were achieved on the combined dataset, which was not attempted or reported previously.

The current analyzes show that even on very short audio sequences, classification is well above guessing, comparable to human precision and ranging in the order of magnitude of 50–60% accuracy, which is still low when relying on it for a single subsequence. However, future work based on the tool could use models designed to combine information over time (as for example pooling over time or hidden Markov Chains) to boost the performance. The SHAP values in Figure 3 offer an empirical basis for evaluating the optimal length of audio segments for emotion recognition models. Higher SHAP values in specific segments suggest that these intervals contain critical emotional information. The consistent presence of such segments across datasets could implie that shorter, information-rich audio clips could be sufficient and potentially more effective for training emotion recognition models. Conversely, segments with lower SHAP values may contribute less to model performance, indicating that longer audio recordings could introduce redundancy or noise. These observations highlight the potential for more efficient model training with carefully selected, shorter audio segments that maximize emotional content. Also, in a time series of emotion classification, some errors may not be as problematic as a miss-classification of a longer, complete audio stream would be. Therefore, it seems plausible that the current approach may allow to generate an emotion time series from an audio stream with sufficient precision.

Comparison with humans

The emotion recognition ability of the models used in this article demonstrated performances comparable to humans, blurring the line between human judgments and model predictions. This suggests that the employed models successfully emulated the human capacity for audio-based emotion recognition in terms of performance. Furthermore, the comparable accuracy between humans and the models implies the involvement of similar mechanisms of pattern recognition.

However, further investigations are required to delve into the intricate workings of the neural network and its alignment with human cognitive processes. This article offers a novel approach to investigate the complexities of audio based human emotion understanding through the application of neural networks. By reverse-engineering such models, valuable insights into the underlying mechanisms and cognitive processes involved in human emotion recognition may be gained. This interdisciplinary research, bridging psychology and computer science, highlights the potential for advancements in automatic emotion recognition and the broad range of applications.

Limitation

The use of actor-performed emotions as the gold standard for developing classification systems may not capture the full range and authenticity of emotions. Actor-performed emotions may not represent the subtler and more authentic emotions often encountered in everyday situations. Given the current state of the models presented, the use of real-life data is questionable due to the databases used. Developing a new dataset that includes a broader range of emotions and different levels of intensity is, therefore, crucial but poses challenges. Heterogeneous datasets containing emotions of varying intensity from different individuals and diverse acoustic qualities may present difficulties in reliably labeling and classifying emotions.

However, this remains the objective, as classifications would ideally be performed on data that closely mirrors reality. In future research enriching the dataset with a broader spectrum of emotions and cultural backgrounds could improve the models' capabilities to recognize a variety of emotional expressions. The exploration of the role played by linguistic differences in emotion recognition could further improve the performance of the models and enhance their practical application.

The influence of linguistic differences on emotion-specific acoustic patterns are another important aspect to consider. Care must be taken to differentiate between patterns that correlate directly with emotions and those influenced by other factors unrelated to emotions. Specializing the classification system in emotion-specific patterns while being resistant to other voice-related information is crucial. Future investigations could delve into the impact of linguistic variations, such as languages and dialects, on the formation of acoustic patterns. By integrating speech recognition into the classification tool, it may be possible to categorize recordings based on language families or cultural linkages. Given the ability to adequately filter acoustic disruptions, such as ambient noise or white noise, the emotion classifier could extend its applications into diverse realms, ranging from everyday interactions to clinical or therapeutic settings.

In these settings, an amalgamation of tools for classifying vocal and facial emotional expressions might offer added benefits. By simultaneously analyzing voice and facial cues, it could pave the way for the creation of adaptive algorithms that generate tailored classification tools, serving both personal and professional needs regarding a wide variety of emotion-related use cases.

Ekman's theory of basic emotions, while easy to interpret, may oversimplify the complexity of human emotions. Considering multidimensional approaches, such as the one proposed by Fontaine et al. (2007), could provide a more nuanced understanding of emotions by defining them across several dimensions. This would accommodate the intricacies and variability of human emotional experiences, allowing for the representation of intermediate emotional states rather than rigid categories like sadness or joy.

In addition, temporal segmentation of audio material into 1.5 s units could lead to forced emotion recognition because it does not capture the natural flow and temporal dynamics of emotions. For example, the CNN design exhibited overfitting, which could be due to the 1.5 s units used. Investigating alternative methods to better capture the temporal dynamics of emotions could potentially enhance the accuracy and generalizability of these models. One method could involve the usage of overlapping windows of clips instead of separate clips.

In the present study we chose a fixed segment length of 1.5 s. The short segment length allows for a continuous classification of human speech and limits overlapping emotions. And the fixed segment length means that continuous human speech would not need to be preprocessed manually into semantically coherent segments. While these short, fixed segments are, hence, necessary for an automatic continuous classification, it is possible that better accuracies can be achieved with longer time segments, as was found in past studies (e.g., Atmaja and Sasou, n.d.[1]). Future studies should investigate whether the use of longer or shorter segments could be advantageous for, in our case, the recognition ability of humans and classifiers. In regard of optimizing the audio file length in terms of maximizing the accuracy of the models, it could be beneficial to include the length as a continuous variable in the model creation pipeline. It is important to emphasize that the present work does not claim to have used the optimal length with the 1.5 s long segments used. In future research, it is recommended to consider employing on-system interpretable systems like SincNet, along with 1D and 2D convolution approaches (Ravanelli and Bengio, 2018; Mayor-Torres et al., 2021), especially for analyzing multimodal signals, such as the audio in this work, as these methods offer promising avenues for enhanced interpretability and analysis.

Enriching the dataset with a broader spectrum of emotions and cultural backgrounds could improve the models' capabilities to recognize a variety of emotional expressions. The exploration of the role played by linguistic differences in emotion recognition could further augment the models' performance. The application-oriented approach demonstrated in this study opens up possibilities for the development of a standalone software application featuring user-friendly interfaces. This application could make the emotion recognition technology more accessible and relevant for real-world implementation.

Conclusion

This article presents a novel approach for classifying emotions using audio data. Through the extraction of features from brief 1.5 s audio segments and the employment of diverse models, we achieved accurate emotion classification across all tested datasets. Our Balanced Accuracies consistently surpassed random guessing. Furthermore, the performance metrics of our DNN and C-DNN models closely mirror human-level accuracy in emotion recognition, showcasing their potential. Nevertheless, the CNN models consistently demonstrated inconsistent results across datasets, indicating limited benefits from employing spectrograms.

In future endeavors, it will be imperative to mitigate overfitting, refine the capture of temporal emotional dynamics, and expand the dataset to encompass a wider range of emotions, cultures, and languages. The creation of a standalone software application equipped with user-friendly interfaces could provide an avenue for the wider application of this emotion recognition technology in myriad settings.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio; http://emodb.bilderbar.info/start.html.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the patients/participants was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

HD: Conceptualization, Methodology, Software, Writing – original draft. LS: Conceptualization, Writing – original draft. TB: Writing – review & editing. TO: Methodology, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

[1] Atmaja, B. T., and Sasou, A. (n.d.). *Multilingual Emotion Share Recognition From Speech by Using Pre-trained Self-supervised Learning Models*. (unpublished).

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Available online at: tensorflow.org (accessed September 5, 2023).

Atmaja, B. T., and Akagi, M. (2020). "Multitask learning and multistage fusion for dimensional audiovisual emotion recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 4482–4486. doi: 10.1109/ICASSP40776.2020.9052916

Burkhardt, F. (2000). *Simulation emotionaler Sprechweise mit Sprachsyntheseverfahren*. (Ph.D. thesis), Technische Universit Berlin, Berlin, Germany.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). "A database of german emotional speech," in *9th European Conference on Speech Communication and Technology* (Marseille: European Language Resources Association), 1517–1520.

Bussmann, H., and Gerstner-Link, C. (2002). "Lexikon der Sprachwissenschaft. Kroener," in *13th International Conference on Digital Audio Effects (DAFX10), Graz, Austria, 2010* (Stuttgart: Alfred Kröner Verlag).

Chambers, J. K. (1973). Canadian raising. *Can. J. Linguist.* 18, 113–135.

Chen, M., He, X., Yang, J., and Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Sign. Process. Lett.* 25, 1440–1444. doi: 10.1109/LSP.2018.2860246

Cheveigna, A. D., and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *Acoust. Soc. Am.* 111:1917. doi: 10.1121/1.1458024

Chourasia, M., Haral, S., Bhatkar, S., and Kulkarni, S. (2021). "Emotion recognition from speech signal using deep learning," in *Lecture Notes on Data Engineering and Communications Technologies, Vol. 57* (Berlin: Springer Science and Business Media Deutschland GmbH), 471–481.

Chung, S. J. (2000). *L'expression et la perception de l'émotion extraite de la parole spontanée: évidences du coréen et de l'anglais*. Doctoral dissertation, Paris.

Davitz, J. R. (1964). *The Communication of Emotional Meaning*. New York, NY: McGraw-Hill.

Driedger, J., and Müller, M. (2014). *Extending Harmonic-Percussive Separation of Audio Signals*. San Francisco, CA: ISMIR.

Dubnov, S. (2004). Generalization of spectral flatness measure for non-gaussian linear processes. *IEEE Sign. Process. Lett.* 11, 698–701. doi: 10.1109/LSP.2004.831663

Ekman, P. (1999). *Basic Emotions. Handbook of Cognition and Emotion* (San Francisco, CA; Chichester: University of California; John Wiley & Sons), 45–60.

Ekman, P., Levenson, R. W., and Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science* 221, 1208–1210.

Ellis, D. (2007). *Chroma Feature Analysis and Synthesis. Resources of Laboratory for the Recognition and Organization of Speech and Audio-LabROSA*, 5. Available online at: https://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/

Fitzgerald, D. (2010). "Harmonic/percussive separation using median filtering," in *Proceedings of the International Conference on Digital Audio Effects (DAFx), Vol. 13* (Graz).

Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychol. Sci.* 18, 1050–1057. doi: 10.1111/j.1467-9280.2007.02024.x

Frick, R. W. (1985). Communicating emotion. The role of prosodic features. *Psychol. Bullet.* 97, 412–429.

Goschke, T., and Dreisbach, G. (2020). "Kognitiv-affektive neurowissenschaft: Emotionale modulation des erinnerns, entscheidens und handelns," in *Klinische Psychologie & Psychotherapie* (Berlin; Heidelberg: Springer), 137–187.

Graham, C., Hamblin, A., and Feldstein, S. (2001). Recognition of emotion in English voices by speakers of Japanese, Spanish and English. *Int. Rev. Appl. Linguist. Lang. Teach.* 39, 19–37. doi: 10.1515/iral.39.1.19

Harte, C., Sandler, M., and Gasser, M. (2006). "Detecting harmonic change in musical audio," in *Proceedings of the ACM International Multimedia Conference and Exhibition* (New York, NY: Association for Computing Machinery), 21–26.

Hung, L. X. (2009). *Detection des emotions dans des ennonces audio multilingues*. (Ph.D. thesis), Institut Polytechnique de Grenoble, France.

Izdebski, K. (2008). *Emotions in the Human Voice, Volume 3: Culture and Perception* San Diego, CA: Plural Publishing Inc.

Jiang, D. N., Lu, L., Zhang, H. J., Tao, J. H., and Cai, L. H. (2002). "Music type classification by spectral contrast feature," in *Proceedings - 2002 IEEE International Conference on Multimedia and Expo, ICME 2002, Vol. 1* (Piscataway, NJ: Institute of Electrical and Electronics Engineers Inc.), 113–116.

Jiang, P., Fu, H., Tao, H., Lei, P., and Zhao, L. (2019). Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition. *IEEE Access* 7, 90368–90377. doi: 10.1109/ACCESS.2019.2927384

Johnson-Laird, P. N., and Oatley, K. (1998). "Basic emotions, rationality, and folk theory," in *Artificial Intelligence and Cognitive Science: Volume 3. Consciousness and Emotion in Cognitive Science: Conceptual and Empirical Issues*. (New York, NY: Toribio & A. Clark).

Jürgens, U. (1979). Vocalization as an emotional indicator a neuroethological study in the squirrel monkey. *Behaviour* 69, 88–117.

Juslin, P. N., and Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol. Bullet.* 129, 770–814. doi: 10.1037/0033-2909.129.5.770

Kim, B., and von Oertzen, T. (2018). Classifiers as a model-free group comparison test. *Behav. Res. Methods* 50, 416–426. doi: 10.3758/s13428-017-0880-z

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., et al. (2010). "Music emotion recognition: a state of the art review," in *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010* (Utrecht), 255–266.

Klapuri, A., and Davy, M. (eds.). (2007). *Signal Processing Methods for Music Transcription*. New York, NY: Springer Science + Business Media LLC.

LeCun, Y., and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *Handb. Brain Theor. Neural Netw.* 3361, 255–258.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2323.

Librosa Development Team (2023). *librosa.magphase*.

Lima, C. F., Castro, S. L., and Scott, S. K. (2013). When voices get emotional: a corpus of nonverbal vocalizations for research on emotion processing. *Behav. Res. Methods* 45, 1234–1245. doi: 10.3758/s13428-013-0324-3

Livingstone, S. R., and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* 13:196391. doi: 10.1371/journal.pone.0196391

Marsh, A. A., Kozak, M. N., and Ambady, N. (2007). Accurate identification of fear facial expressions predicts prosocial behavior. *Emotion* 7, 239–251. doi: 10.1037/1528-3542.7.2.239

Mauch, M., and Dixon, S. (2014). PYIN: a fundamental frequency estimator using probabilistic threshold distributions. *IEEE Expl.* 2014:6853678. doi: 10.1109/ICASSP.2014.6853678

Mayor-Torres, J. M., Ravanelli, M., Medina-DeVilliers, S. E., Lerner, M. D., and Riccardi, G. (2021). "Interpretable sincnet-based deep learning for emotion recognition from EEG brain activity," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (Mexico: IEEE), 412–415.

Miller, P. W. (1981). Silent messages. *Childh. Educ.* 58, 20–24.

Moors, A., Ellsworth, P. C., Scherer, K. R., and Frijda, N. H. (2013). Appraisal theories of emotion: state of the art and future development. *Emot. Rev.* 5, 119–124. doi: 10.1177/1754073912468165

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *Am. Natural.* 111, 855–869.

Morton, J. B., and Trehub, S. E. (2001). Children's understanding of emotion in speech. *Child Dev.* 72, 834–843. doi: 10.1111/1467-8624.00318

Mustaqeem and Kwon, S. (2019). A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors* 20:183. doi: 10.3390/s20010183

Mustaqeem and Kwon, S. (2021). Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network. *Int. J. Intell. Syst.* 36, 5116–5135. doi: 10.1002/int.22505

Mustaqeem, Sajjad, M., and Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLDTM. *IEEE Access* 8, 79861–79875. doi: 10.1109/ACCESS.2020.2990405

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Machine Learn. Res.* 12, 2825–2830. Available online at: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Ravanelli, M., and Bengio, Y. (2018). Interpretable convolutional filters with sincnet. *arXiv preprint arXiv:1811.09725.* doi: 10.48550/arXiv.1811.09725

Sandhya, P., Spoorthy, V., Koolagudi, S. G., and Sobhana, N. V. (2020). "Spectral features for emotional speaker recognition," in *Proceedings of 2020 3rd International Conference on Advances in Electronics, Computers and Communications, ICAECC 2020*. Bengaluru: Institute of Electrical and Electronics Engineers Inc.

Sato, N., and Obuchi, Y. (2007). Emotion recognition using mel-frequency cepstral coefficients. *J. Nat. Lang. Process.* 14, 83–96. doi: 10.5715/jnlp.14.4_83

Scherer, K. R. (1979). *Nonlinguistic Vocal Indicators of Emotion and Psychopathology*. (Boston, MA: Springer), 493–529.

Scherer, K. R. (1985). Vocal affect signaling: a comparative approach. *Adv. Study Behav.* 15, 189–244.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Soc. Sci. Inform.* 44, 695–729. doi: 10.1177/0539018405058216

Smith, J. O. (2011). *Spectral Audio Signal Processing*. W3K. Available online at: https://cir.nii.ac.jp/crid/1130282272703449216

Stresemann, L. (2021). *AVECT: Automatic Vocal Emotion Classification Tool*. (Ph.D. thesis), Universität der Bundeswehr München.

Trojan, F., Tembrock, G., and Schendl, H. (1975). *Biophonetik*. Available online at: https://catalog.loc.gov/vwebv/search?searchCode=LCCN&searchArg=75516032&searchType=1&permalink=y

Xiao, Z., Dellandrea, E., Dou, W., and Chen, L. (2010). Multi-stage classification of emotional speech motivated by a dimensional emotion model. *Multimedia Tools Appl.* 46, 119–145. doi: 10.1007/s11042-009-0319-3

Appendix

TABLE A1 **Overview of hyperparameters used for model creation.**

Hyperparameter	Options	Description
Number of layer	2 to 8	Describes the depth of the model
Number of neurons	80 to 400	Describes the size of the layers
Activation function	relu	$f(x) = \max(0, x)$
	elu	$f(x) = \begin{cases} \alpha(e^{x} - 1) & \text{for } x \leq 0 \\ x & \text{for } x > 0 \end{cases}$
	sigmoid	$f(x) = \frac{1}{1+\exp(-x)}$
	tanh	$f(x) = \frac{e^{x}-e^{-}x}{e^{x}+e^{-}x}$
Optimization function	SGD	Optimized using a random training example
	RMSprop	Optimized using a adaptive Learning rate
	Adam	Combination of momentum optimization and adaptive learning rate
Error function	Categorical cross entropy loss	$-\sum_{i}^{n} y_i \log(\hat{y}_i)$
Learning rate	0.01 to 0.0000001	The step size that the model takes toward minimizing the cost function

This table provides a non-exhaustive overview of the different hyperparameters used in model creation along with their options and descriptions.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 27 March 2024
DOI 10.3389/fpsyg.2024.1363174

OPEN ACCESS

EDITED BY
Simone Belli,
Complutense University of Madrid, Spain

REVIEWED BY
Meisam Dastani,
Gonabad University of Medical Sciences, Iran
Vasile Daniel Pavaloaia,
Alexandru Ioan Cuza University, Romania

*CORRESPONDENCE
Ayhan Dolunay
✉ ayhan.dolunay@neu.edu.tr

RECEIVED 29 December 2023
ACCEPTED 26 February 2024
PUBLISHED 27 March 2024

CITATION
Dolunay A and Temel AC (2024) The relationship between personal and professional goals and emotional state in academia: a study on unethical use of artificial intelligence.
Front. Psychol. 15:1363174.
doi: 10.3389/fpsyg.2024.1363174

The relationship between personal and professional goals and emotional state in academia: a study on unethical use of artificial intelligence

Ayhan Dolunay[1]* and Ahmet C. Temel[2]

[1]Faculty of Communication, Grand Library, Near East University, Nicosia, Cyprus, [2]Grand Library, University of Kyrenia, Kyrenia, Cyprus

Artificial Intelligence (AI) is a concept that has been a subfield of computer science since the 1950s. In recent years, with its growing development power, AI technologies have made significant progress and are now being used in many fields. Like in all areas, the use of AI technologies in academia has provided convenience to academics while also bringing ethical debates. In the literature part of the study, concepts such as AI, academia, academics and academic progress, ethics, ethical theories, academic ethics, and emotional states have been thoroughly examined and defined. In this study, starting from AI and scientific ethics, ethical issues arising from emotional states in academic research have been identified, and concrete solutions to these ethical issues have been proposed. The aim is to discuss the views of academics in order to determine what types of scientific ethical violations and prevention methods are involved. In this context, the semi-structured interview technique, which is one of the qualitative research methods, was preferred as the method. In the study, in-depth semi-structured interviews were conducted with 4 ethics experts and 4 psychology experts selected through snowball sampling technique. The data obtained through semi-structured in-depth interviews will be analyzed using content analysis. Within the context of the literature review and interviews: Ethics is based on the foundation of acting correctly. In this context, scientific ethics can be summarized as acting truthfully and honestly, not distorting data, and not trying to progress unfairly. The use of AI in academia is becoming increasingly widespread. From a positive perspective, this usage significantly contributes to making studies more practical. However, it can lead to problems such as unfair authorship, devaluation of human authorship, and incorrect data. The connection between academics' professional advancement goals and emotional states becomes prominent in this context. The potential of AI to facilitate progression can lead to unethical use. To prevent such situations, it is recommended to organize training sessions to increase professional awareness, internalize ethics personally, establish ethical committees specific to the field of AI, conduct more effective audits by academic publication and promotion committees, and implement specific regulations for AI. Finally, for future academic studies, it is suggested that the usage of AI in academic research be measured and evaluated by ethics experts. For psychologists, conducting surveys with academics to explore how they use AI in the context of their emotional states and professional advancement goals is recommended.

KEYWORDS

academia, ethics, academics, emotions, artificial intelligence

1 Introduction

Digitalization, as in every field, has led to the progressive development and transformation of AI. Similar to many other fields, AI has undergone significant changes from its emergence in 1956 (Kokina and Davenport, 2017) to the present day (See Chiu et al., 2023). This autonomy and development, as all aspects of life, have various implications in academia. These effects have sparked numerous debates, both positive and negative perspectives.

Many tasks previously performed by humans can now be carried out by machines and algorithms. For example, tasks such as article segmentation, analysis, and data processing can now be done more quickly and effectively with the help of AI (Mijwil et al., 2023). As a result, there has been a transformation process in the academic world.

This transformation process directly affects the professional progression and emotional state of academics. On one hand, the use of AI allows for faster and more efficient work, but on the other hand, these technological advances have caused academics to question their roles and abilities and redefine themselves. The tasks performed by AI have prompted academics to question the topics they have previously worked on and have led to changes in research areas (Altıntop, 2023).

On the other hand, with the increasing use of AI, various debates have emerged in the academic world. While AI provides great convenience to academics in areas such as topic suggestions, editing sections, data analysis, it also raises concerns in areas such as knowledge sharing, the threat of eliminating human authorship, unethical behavior, misinformation, creativity, and human-specific skills. Concerns such as the replacement of humans and the decrease in the human factor have sparked debates among academics (Crompton and Burke, 2023).

This study specifically examines the connection between the use of AI and the professional progression and emotional state of academics. Taking into account the advantages and disadvantages brought about by the use of AI, the study aims to analyze the impact on academics' career development and emotional state. This study is an important step toward better understanding the changes brought about by the use of AI in the academic world and discussing possible future impacts.

And specifically conducts research and discussion on the connection between the use of AI, the professional advancement of academics, and their emotional states.

2 Literature review

2.1 Basic concepts

2.1.1 Artificial intelligence

AI refers to the ability of computers or computer-assisted machines to perform high-level logical processes that humans are capable of. These abilities include finding solutions, understanding, deriving meaning, generalizing, and learning (Öztürk and Şahin, 2018).

AI is a concept used to give computer systems human-like capabilities. AI enables computer systems to analyze tasks similar to human intelligence, including analysis, learning, problem solving, and decision making (Öztürk and Şahin, 2018).

The history of AI is quite extensive. The Dartmouth Conference in 1956 is considered the birthplace of AI. Since then, AI has rapidly developed worldwide. The concept of AI was popularized in a letter proposed by John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon during the Dartmouth Conference in 1956. Although John McCarthy is mentioned as the creator of this concept, AI was evaluated as an important step in the birth of AI in the offer letter (Arslan, 2020).

In the early stages, AI technology was limited to large computer systems and specialized software. However, nowadays, AI is being used in various fields such as mobile devices, smart home systems, automobiles, healthcare, and education (Öztürk and Şahin, 2018).

The applications of AI are vast. AI can be used in e-commerce websites to track customer behavior, in the financial sector for credit risk analysis, in traffic management, in the healthcare sector for diagnosis and treatment planning, and in automated factories, among many other fields. Furthermore, AI technology also makes it easier for people in their daily lives. You can use voice commands to quickly search on smartphones, create personalized music playlists, control devices at home using voice commands, and even have your emails written for you (Arslan, 2020).

While AI is used in many different fields, it also has significant effects in academia. At this point, it would be appropriate to briefly define the concepts of academia and academics.

2.1.2 Academia, academics, and academic advancement

The origin of the word "academia" is attributed to Plato's school named "Akademia" near Athens. Today, the word "academia" is used interchangeably with the word "university," which is defined as an educational institution that is centered around science and where knowledge is produced and disseminated. However, there are differences in meaning and function between Plato's Akademia and today's universities. The emergence of the first universities took place after the 11th century, and over time, the functions and expectations of universities have diversified (Akcan et al., 2018).

Universities have undertaken various roles throughout history (Aydin, 2016: 14–20). The prominent ones among these roles are: generating knowledge, disseminating knowledge, providing professional education, imparting general culture, serving the community, and finally, being an actor in the global economy. Gasset (2014, 52) argues that universities have three main tasks: to ensure the cultivation of any individual, to provide the necessary knowledge and experience to perform any profession, and to train researchers (Akcan et al., 2018).

Academia forms the foundation of the concepts of academician/academics. In the process that has survived from the past to the present, this term began to be used for a certain career and the individuals who hold this career (Gürkan, 2018).

An academician is someone who has received undergraduate education in a discipline, gained expertise by pursuing a graduate education in the same or a different discipline, and works at a university. Academic career is a type of career that offers a wide range of opportunities, goes beyond a specific job, and represents not only work but also a way of life and thinking (Gürkan, 2018).

The academic advancements or promotions of academics are also important in terms of the subject matter. There are three evaluation methods commonly used and considered appropriate for academic promotions (Demir et al., 2017): *Academic publications and citations received, Practices in education and teaching, University and community service.*

For example, in promotions to the rank of associate professor in Turkey, oral examinations are also used as a criterion in addition to these. Generally, globally, measurement and practice are widespread, with the highest emphasis on the first category. Academic advancements require a meticulous examination and adherence to ethical rules (Demir et al., 2017).

The evaluation method commonly used in academic advancements is the academic publications of academics and the citations they receive. In this method, the number, quality, and impact of articles published by an academician, as well as the number of citations.

2.2 Ethical theories and their adaptation to academic ethics

Setting aside the definition and boundaries of the concept of ethics, briefly exploring the approaches of philosophers to ethics throughout the ages will contribute significantly to explaining the current understanding of ethics and academic ethics. This is because ethics, being a concept that has existed for centuries, still holds great importance.

Socrates, who lived in Athens between 461 and 399 BCE, endeavored to educate the people of Athens on ethical matters. He emphasized the significance of knowledge in making ethical decisions, pointing out that ignorance is one of the main causes of wrong decisions. Applied to academics, this implies that scholars can make ethical decisions only when they are knowledgeable about the subject at hand, maintaining its validity in contemporary times.

According to Plato, a student of Socrates, virtues such as moderation, courage, and wisdom come together to create the highest virtue, forming justice. Plato's concept of justice is broader than the contemporary understanding, signifying a moral-good life seen as the ultimate good (Peck and Reel, 2013: 9; Dolunay, 2018). In other words, to achieve a good life, one must obtain a morally good life. This attainment, for academics, can contribute to the awareness of society and individuals.

Aristotle, significantly influenced by Socrates and Plato, believed that ethical decision-making is a skill (techne) and that ethical behavior cannot be a precise science because there is no formula that fits every situation. Aristotle also advocated avoiding extremes. He viewed virtue as a middle ground between excess and deficiency. Aristotle saw the acquisition of the right character through education as essential for making the right choices. Learning from books, intellectual virtues gained through reading ethical rules, is another aspect of Aristotle's philosophy (Peck and Reel, 2013: 10; Dolunay, 2018).

Aristotle argued that ethical virtues are learned through actions and must be acquired as habits. This doctrine requires possessing the right character for ethical behavior. In the context of today, academics using data obtained through 'unethical' means in their academic research, exceeding ethical boundaries in the application of AI, can be considered an excess. In such situations, finding a middle ground between excess and deficiency by behaving virtuously becomes crucial. The 18th-century philosophers, Bentham defined the principle (basic utilitarianism and the utility theory) currently considered a classic approach in terms of pleasure and pain instead of benefits and harms (Peck and Reel, 2013: 13; Dolunay, 2018). When applied to academics, this theory highlights the necessity for scholars to balance individual progress in their academic field with the respect and reputation of academia. Individual progress is important for the development of the field, but academics must achieve this within legal and ethical boundaries.

In the 20th century, Ross believed in *prima facie* duties, including keeping promises (fidelity), showing gratitude for good, being fair, improving the lives of others (beneficence), avoiding harm, making amends when necessary (reparation), and self-improvement. Ross did not consider these duties as the only ones, allowing for the list to be expanded. In some ethical dilemmas, multiple duties may apply. In such cases, individuals must decide which duty takes precedence for that particular situation (Peck and Reel, 2013: 16; Dolunay, 2018).

Contemporary philosopher Rawls, a Harvard professor, created a concept of justice that many students find useful in ethical decision-making. In this context, a person should ignore their own position, placing themselves behind a veil of ignorance, and make decisions (Peck and Reel, 2013, p 17; Dolunay, 2018). Applied to academics, this theory suggests that an academic, when conducting research or publishing, should consider the potential harm to individuals or groups and the impact on academia and the requesting institution without being aware of their hierarchical position.

In conclusion, exploring the ethical perspectives of philosophers across different eras provides valuable insights into the current understanding of ethics and its application in academia.

2.3 Academic ethics and emotions of academics relations

The word "ethics" originates from the French word "éthique," which in turn comes from the Old Greek word "ethios," meaning character and moral. This term carries the meaning related to morality. The word "ethios" is derived from the Old Greek word "ethos," which encompasses custom, morality, tradition, and manners (Dolunay, 2018: 26).

According to Pieper, "ethics is not only a theoretical scientific concept but also something that can be practically realized" (Uzun, 2007; Dolunay and Kasap, 2018). In other words, ethics does not have meaning on its own but gains significance when associated with something, such as academic ethics.

In scientific research, ethics refers to the moral principles and norms that scientists must adhere to in the research and publication processes. Scientific ethics aims to ensure the accuracy, reliability, and societal benefit of science. Adhering to ethical rules in scientific research enhances the reputation of both scientists and the scientific field (Yördem and Şeker, 2018).

Ethical rules in scientific research include: Truthfulness, diligence, transparency, impartiality, social benefit, education, appreciation, and avoiding 'ethical violations, improper citations, fabrication, falsification, duplication, fragmentation, unjust authorship' (Resnik, 2012).

On the other hand, all these ethical values, especially in the context of the academic advancement goals of academics, need to be carefully considered. While the significance of emotions and advancement goals has been discussed in various studies, it is crucial to avoid unethical approaches driven by emotions and advancement goals to achieve success more rapidly. Otherwise, with the influence of advancement goals and aspirations (negative emotional states), unethical situations such as unjust progress, persistent seeking of recognition and the spotlight in academia, desire to see one's name frequently in academic publications (the Hollywood effect) can emerge (Ercan et al., 2021).

In order to be successful in the academic field, it is generally a vital goal for academics. The motivation and ambition necessary to achieve success can encourage academics to become better researchers, writers, or academics. However, it is a fact that an academic who cannot control their emotions may resort to unethical behavior for achieving success (Maya, 2013).

Ethics in the academic field is based on principles of honesty, impartiality, and respect. When conducting research, evaluating students, or preparing publications, academics should be objective and focus on universal knowledge and principles of justice instead of personal interests. However, an excessive desire for success can lead an academic to deviate from ethical principles (Tunç, 2007).

An academic lacking emotional control may try any means to compete with colleagues. They may use another academic's work without permission, manipulate results, or work unfairly to gain an advantage over other researchers. Such behaviors can undermine trust in the academic field, affect the work of other researchers, and harm the scientific community (Maya, 2013).

These unethical behaviors prevent everyone in the academic field from feeling safe and in a fair environment. Respecting the value of everyone's academic work is important to allow individuals to freely express their ideas and progress objectively (Maya, 2013).

In conclusion, it is normal to strive for success in the academic field, but an academic who cannot control their emotions may resort to unethical behavior. Academics need to be conscious of emotional control for the development of the academic community and equal opportunities for individuals. Upholding principles of objectivity, honesty, respect, and justice is important to avoid unethical behavior (Tunç, 2007).

At this point, especially with the involvement of the use of AI, the situation becomes more complex.

One of the most important developments in human history can be considered the rise of AI. While this technology has had a significant impact in various sectors, it has also caused important transformations in the academic field. However, although AI is a tool that supports and enhances people's work, it can sometimes be subjected to unethical uses (Ülman, 2006).

An academic becoming excessively ambitious and losing control of their emotions in order to achieve success can also lead to the unethical use of AI. This situation presents behavior that contradicts ethical rules and human values in the scientific world. Considering that academics have a mission to produce knowledge, explore, and enhance the well-being of society, using AI unethically would be an approach that undermines this mission (Ülman, 2006).

An academic using AI negatively in order to achieve their goals is also contrary to the concept of scientific ethics. Science is a discipline that promotes objectivity, impartiality, and freedom of thought. Therefore, using a tool like AI, which supports scientific research, in the shadow of personal ambitions and emotions, both damages trust in science and misdirects knowledge construction (Altıntop, 2023).

Another unethical aspect of using AI in the shadow of an academic's emotional state is the misuse of information. AI technologies, with their ability to perform big data analysis and make predictions, enable academics to quickly attain important results. However, an academic taking advantage of these rapid results by presenting fake data or manipulating results can cause great harm to the scientific community and society (Altıntop, 2023).

In this context, the pursuit of success in the academic field by academics losing control of their emotions and using AI unethically is an unacceptable situation from ethical, scientific, and societal perspectives. Academics must firmly adhere to ethical values while fulfilling their scientific responsibilities. Academics who combine the advantages provided by AI with ethical and human values will contribute to future scientific advancements and societal benefit (Altıntop, 2023).

In this framework, it would be appropriate to discuss academia and AI separately under a separate heading and briefly touch upon their positive and negative effects.

2.4 Artificial intelligence and academia

AI refers to the ability of computers or computer-supported machines to perform high-level logical processes that are typically associated with human capabilities. These skills include finding solutions, understanding, deriving meaning, generalizing, and learning (Muthukrishnan et al., 2020).

The term AI is used to describe the concept of giving computer systems human-like features. AI empowers computer systems to analyze, learn, solve problems, and make decisions in a manner similar to human intelligence (Muthukrishnan et al., 2020).

The history of AI is quite extensive. The Dartmouth Conference in 1956 is considered the birthplace of AI, where its foundations were laid. Since then, AI has rapidly developed worldwide. The concept of AI first emerged in a proposal letter presented at the Dartmouth Conference in 1956 by John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. While John McCarthy is remembered as the creator of this concept, the proposal letter is considered a significant step in the birth of AI (Kokina and Davenport, 2017).

The use of AI is prevalent in various fields, including academia, serving various purposes. AI possesses capabilities such as teaching, structuring articles, conducting research and data analysis, and handling large-scale data examination and analysis, among other features (See Chiu et al., 2023).

2.4.1 Positive perspective

AI has numerous positive aspects, and a few of them are outlined below:

Teaching at the university level: AI possesses the capability to instruct courses that require expertise in a specific subject. For example, Assistant Professor Dux at Near East University is an AI instructor. Through AI, students can more effectively access the courses they need (Mijwil et al., 2023).

Article structuring: AI can automatically divide a chosen topic into sections. By identifying key words or topics in texts, AI can also suggest titles and section headings. This allows for quick structuring of articles with less time investment (Mijwil et al., 2023).

Conducting analyses: AI has the ability to perform rapid and precise analyses on large datasets. For instance, an AI system can analyze uploaded text, images, or audio. These analyses assist in determining better strategies and making informed decisions (Mijwil et al., 2023).

Language translation: AI can translate text into multiple languages, facilitating communication among individuals who speak different languages. Additionally, it enables understanding of international articles or books in one's own language (Lund et al., 2023).

2.4.2 Negative perspective

One of the primary negative features of AI is the potential for unfair content generation. AI algorithms can generate new content by analyzing large amounts of data. However, there may be limitations on the accuracy and sensitivity of this content. AI can allow people to disseminate misinformation or produce inaccurate content (Sariyasa and Monika, 2023).

Another concern related to the use of AI is its potential for unethical behavior. People can use AI for wrongful purposes. For example, an individual or group that does not put in effort may use AI to generate content effortlessly and achieve success as a result (Thunström, 2022).

The advancement of AI poses a risk of eliminating human writing. AI algorithms can produce complex writings and reports, taking over tasks that many people currently perform. This situation could lead to unemployment in this field (Jabotinsky and Sarel, 2022).

Furthermore, AI can pose a threat to unfair promotion and academic progress. For instance, an automatically generated article or thesis that appears high-quality can be produced using AI. In such cases, many individuals may present these articles and theses as original works, leading to undeserved academic success. In summary, some of the negative features of AI include unfair content generation, unethical behavior, the risk of eliminating human writing, and the possibility of unfair progress and academic advancement. It is important to consider these concerns and regulate the development of AI (Thunström, 2022).

For example, there are significant debates in the literature about the negative and destructive effects of using AI in the field of education (Păvăloaia and Necula, 2023). These can be listed as technology addiction in education, the problem of determining responsibility in the event of a potential error, concerns about individuals losing their jobs, and issues related to data collection and analysis.

Technology Addiction: Individuals can become excessively dependent on AI-supported educational tools, deviate from traditional learning methods, and become detached from real-world interactions (Păvăloaia and Necula, 2023).

Responsibility Issue: When learning content or decisions provided by AI are incorrect, determining who is responsible can become uncertain. This situation may lead to disagreements on responsibility between educational institutions and technology providers (Sáiz-Manzanares et al., 2022).

Risk of Unemployment: If certain traditional teaching roles are taken over by AI and automation, teachers and other education professionals may face the risk of unemployment.

Inequality and Discrimination: AI algorithms can reflect biases and deepen inequalities in education. For example, equal opportunities may not be provided to students based on factors such as their ethnicity, gender, or socioeconomic status (Sáiz-Manzanares et al., 2022).

In order to address these issues, it is important to carefully establish governance, regulation, and ethical standards in the development and implementation of AI-supported education systems. Additionally, awareness of the risks associated with the use of AI technologies in education and continuous efforts to mitigate these risks are necessary.

3 Research

3.1 Method

In the study, the aim is to discuss the opinions of ethics experts and psychology specialists. In this context, the method chosen is the semi-structured interview technique, which is one of the qualitative research methods.

Interviews are used as a professional technique or auxiliary tool in many social science fields such as journalism, law, and medicine (Kahn, 1983; Tekin, 2006: 101). An extensively used data collection technique in qualitative research, interviews provide the interviewed individuals with the opportunity to express themselves directly, while also allowing the researcher to observe the interviewee comprehensively (McCracken, 1988: 9; Tekin, 2006: 102).

The interviewed individuals were asked questions covering all dimensions of the research topic, and detailed answers were obtained; it is a technique that enables the direct collection of information (Johshon, 2002: 106; Tekin, 2006: 102). Interviews can be categorized as unstructured, semi-structured, and structured (Punch, 2005: 166; Tekin, 2006: 104). Semi-structured interviews use predetermined questions, making them more limited compared to unstructured interviews, but it is possible to ask spontaneous questions and elaborate on targeted data/responses based on the course of the interview.

The data was collected through interview forms prepared by the authors of the study, containing 5 questions for psychology experts and 7 questions for ethics experts. Interviews were conducted between September 2023 and November 2023 over a period of 2 months. The interview forms were delivered to participants online, and they were asked to provide written answers to ensure no missing or lost responses. The collected data was archived in an online cloud database.

3.2 Sampling

The study involved semi-structured in-depth interviews with four ethics experts and four psychology specialists selected through the snowball sampling technique.

The presence of ethics experts in the interview group is primarily due to the need to discuss the ethical implications of AI usage. The development of AI technology has brought forth numerous ethical issues, necessitating the need to address or manage these issues. Therefore, by including ethics experts in the interview group, the aim is to discuss the emerging problems by obtaining views and recommendations on the use of AI and ethics in academia.

On the other hand, the inclusion of psychology experts in the sample is primarily aimed at investigating and understanding the potential emotional effects that may arise during the use of AI by academics. Unethical uses of AI are a significant factor in the emotional effects it may have on individuals in their daily lives and/or professional careers. Understanding and addressing such emotional states requires the insights and recommendations of psychology experts.

Thus, while ethics experts provide opinions and recommendations on the ethical use of AI, psychology experts will enrich the study with their responses on the emotional effects and mood states related to AI. Their collaboration enables a more comprehensive assessment of the ethical and emotional dimensions of AI, contributing to evaluating the potential consequences of unethical uses and fostering interdisciplinary work.

The snowball sampling technique was employed, selecting individuals based on their expertise in their respective fields and a minimum of 5 years of professional experience. Additionally, another important criterion was the selection of academics who have studies or knowledge in this area. Therefore, professionals with both professional experience in the field and knowledge through studies or research in the relevant field were selected for the study.

The selected individuals were asked to recommend others who meet these criteria. Among the experts in the field, the criterion of a minimum of 5 years of professional experience shaped the interview group, ranging from 5 to a maximum of 25 years of professional experience. The snowball sampling technique is a method that involves selecting a reference person related to the subject of the study and reaching other individuals through recommendations. This method is iterative, and participants guide researchers, contributing to the growth of the sample. Therefore, it is known as the "snowball effect" (Biernacki and Waldorf, 1981).

3.3 Analyses

The data obtained from semi-structured in-depth interviews were analyzed through content analysis. Content analysis is a research technique where valid interpretations extracted from the text are revealed through consecutive processes (Weber, 1990: 9; Koçak and Arun, 2013: 22). Depending on the context of a specific study, detailed coding may or may not be required (Yıldırım and Şimşek, 2008: 233; Karataş, 2017: 80).

In this context, due to the nature of the study, there was no need for intricate coding and theme formation. The themes and codes are as follows:

In the analyzes under the specified themes and codes (Table 1), 20% of the direct opinions of the interview group were included. The names of the participants are given in codes as P1, P2, P3, etc.

4 Terms of 'ethics' and academic ethics

4.1 Terms of 'ethics'

Ethics is a concept that is difficult to define. Generally, it can be defined as a research discipline where moral situations are described, observation tools are developed; criteria are constructed based on what is good and bad or what is right and wrong, and a critical demand where they are validated (Moressi, 2006: 23; Girgin, 2000: 144).

Similarly to this definition, the interview group has also provided a response in line with the prevailing consensus in the literature.

The interview group has collectively defined the concept of ethics as acting in accordance with the correct principles and behaving in accordance with professional fundamental principles:

P2: "*Determining what the correct way to act could be. In other words, how should one act?*"

P4: *"I define the concept of ethics as the identification of individual, professional, institutional, and societal values, and the use of these identified values as a criterion for evaluating human behavior."*

4.2 Academic ethics

In scientific research, ethics refers to the moral principles and norms that scientists must adhere to in the research and publication processes. Scientific ethics aims to ensure the accuracy, reliability, and societal benefit of science. Adhering to ethical rules in scientific research enhances the reputation of both scientists and the scientific field (Yördem and Şeker, 2018).

The interview group also made a similar definition. Scientific ethics is defined in accordance with the concept of ethics as being focused on acting correctly, behaving honestly and fairly, working within boundaries that are beneficial and respectful to society and nature.

P1: *"All scientific research is conducted with the aim of finding truth, discovering new things, and finding solutions to observed problems. And while all of this is done, benefiting both the field and society, and humanity are fundamental goals; therefore, scientific ethics are indispensable."*

P3: *"I evaluate scientific ethics in two ways. Firstly, individuals conducting scientific research should behave sensitively towards the environment and living beings related to the subject they are working on. In scientific research, to reach a conclusion, one should avoid behaviors that could harm the environment or cause physical or mental harm to living beings. Secondly, individuals conducting scientific research should not use any information or documents derived from previously conducted sources or sources of inspiration without citing references."*

TABLE 1 Themes and codes.

Themes				
Codes	Terms of 'ethics' and academic ethics	AI in academy and positive–negative effects of using AI in academy	Relation between emotions of academics, AI and academics professional goals	ethical problems raised by AI and solution suggestions in the context of scientific ethics
	Terms of 'Ethics'	AI in Academy	Emotional Situation	Ethical Problems
	Academic Ethics	Positive and Negative Effects	*Unfair Academic Professions and AI*	Solution Suggestions

5 AI in academy and positive–negative effects of using Ai in academy

5.1 AI in academy

The use of AI is prevalent in various fields, including academia, serving various purposes (Chiu et al., 2023).

The interview group has expressed a common view that AI and AI technologies are used in the academic field. They have emphasized that the evolving and changing technology influences academia and that AI is utilized for both structuring and writing in academic research:

P2: *"I believe the application for AI in academic fields can be a way for a student, instructor, or researcher to begin their research…."*

P3: *"AI has recently contributed to both students' and academics' easy access to information, while also taking on an educational role with various developed AI modules."*

5.2 Positive and negative effects

In academia, AI possesses capabilities such as teaching, structuring articles, conducting research and data analysis, and handling large-scale data examination and analysis, among other features (See Chiu et al., 2023). On the other hand, AI can allow people to disseminate misinformation or produce inaccurate content (Sariyasa and Monika, 2023). Same time, an individual or group that does not put in effort may use AI to generate content effortlessly and achieve success as a result (Thunström, 2022) AI algorithms can produce complex writings and reports, taking over tasks that many people currently perform (Jabotinsky and Sarel, 2022).

In accordance with the positive and negative aspects given above in the literature, the interview group has identified both positive and negative aspects of the use of AI in academia.

In this context, the positive aspects are as follows: time savings, easy access to resources, support in text writing, contribution to structuring and analysis, and assistance in creating visuals and tables:

P1: *"While AI provides us with many advantages like this, especially when conducting research, it saves time, facilitates cost, and perhaps brings us together with resources that may be difficult to reach. Therefore, I view its use in academic studies positively because it has many advantages in* various *positive aspects."*

P4: *"Its positive aspects can assist scientists in finding academic sources in academic studies. It can help in applications that academics may not easily accomplish, such as data visualization."*

On the other hand, negative aspects include: unfair authorship and/or devaluation of human authorship, the possibility of inaccurate data, plagiarism, and the potential for achieving results with minimal effort:

P1: *"As researchers, we should reevaluate the information provided by AI, strive to reproduce it, and look beyond the framework it presents to us. Of course, in addition to this, we should support the given information with our own ideas… Otherwise, the role of the researcher may deviate, various ethical issues may arise, the researcher's image may be damaged, which may not be limited to the researcher alone but may also lead to questioning the discipline and credibility of the relevant field."*

P4: *"As negative aspects, it can produce texts instead of academics. It can lead to plagiarism, and detecting it may not be easy."*

6 Relation between emotions of academics, AI and academics professional goals

6.1 Emotional situation

Various studies in the literature indicate that there is a connection between mood states and the behaviors of academics from different perspectives (Maya, 2013). Similarly, the interview group holds the opinion that the mood states of academics are linked to their professional behaviors. However, within the focus of the study's context, the interview group was asked for their opinions on whether academic advancement goals are also linked to mood states.

Questions about emotional states were directed specifically to psychology experts within the context of their expertise. The relevant group is of the opinion that the emotional states of academics are generally connected to their professional advancement goals:P5: "Ethics is a moral understanding in my opinion. Therefore, even if there are written rules, whether to comply with them or not is still within one's personal discretion. Therefore, unfortunately, it is indeed possible to deviate from ethical rules within the framework of personal ambition and goals."

P5: "Ethics is a moral understanding in my opinion. Therefore, even if there are written rules, whether to comply with them or not is still within one's personal discretion. Therefore, unfortunately, it is indeed possible to deviate from ethical rules within the framework of personal ambition and goals."

6.2 Unfair academic professions and AI

The interview group (psychology experts), expressing that emotional states and academic progress are interconnected, predominantly believes that, simultaneously, the influence of emotional states may lead to unfair use of AI in the context of academics' career advancement goals. However, they also consider that some AI applications are not yet as competent in this regard:

P8: "In the context of academics' ambitions and advancement goals, unfair or unethical use of AI may be possible."

7 Ethical problems raised by AI and solution suggestions in the context of scientific ethics

7.1 Ethical problems

The ethical issues arising from the use of AI in academia include "the distortion and/or inaccuracy of data, unfair authorship, the formation of plagiarism, and reaching a correct or incorrect result without exerting effort."

In this context, especially in the context of emotional states and career advancement goals, academics' unjust use of AI driven by these motives can pose a significant ethical problem:

P3: "The ethical issue arising from the use of AI in academic studies may occur when researchers present information derived from AI in their studies as if they had produced it themselves, rather than generating subjective knowledge."

P4: "The most serious ethical issue is when an academic has their academic work done by AI. AI can easily generate data and interpret it into an article. Additionally, it can generate imaginary citations."

7.2 Solution suggestions

Those who provided recommendations against the ethical problems that may arise from unfair use of AI in the interview group have put forward the following solution proposals:

Implementation of professional awareness and training activities,

Individual internalization of ethical values and understanding that unfair progression is not appropriate in this context,

More careful evaluations by publication and/or academic promotion committees, utilizing more comprehensive technological control practices specific to the field,

Development/updates of ethical principles/rules in the context of AI,

Establishment of ethical committees specific to the use of AI.

P1: "In this regard, scientific education programs should be organized, boards should conduct more active monitoring, and regulations need to be developed.."

P5: "...the peer (science) review board should be more meticulous in examining studies, and if they detect the use of AI, researchers should face more serious sanctions."

8 Findings and discussion

This study examines the relationship between the unethical use of AI in academia and the personal and professional goals, as well as the emotional states of academics. Findings obtained through interviews indicate various significant results.

Firstly, the findings of our research emphasize that scientific ethics is based on proper and honest conduct. In this context, scientific ethics involves acting honestly and accurately without distorting data and advancing with unjust motives. Ethics is based on the foundation of acting correctly. In this context, scientific ethics can be summarized as acting truthfully and honestly, not distorting data, and not trying to progress unfairly.

However, the increasingly widespread use of AI in academia poses new challenges to these ethical standards. As findings, it has been determined that artificial intelligence provides speed and practicality in academic studies. In particular, providing topic suggestions, determining the main sections of the studies, and contributing to analyzes provide significant convenience. On the other hand, it has been determined that the use of artificial intelligence in academic studies may also lead to negative situations such as reducing the value of human authorship, causing unfair authorship, and providing inaccurate data.

These findings are consistent with the views in the literature. While AI makes research more practical (Mijwil et al., 2023), it can also lead to issues such as unfair authorship, diminished value of human authorship, and incorrect data (Sariyasa and Monika, 2023).

On the other hand the research highlights the importance of the connection between academics' professional advancement goals and emotional states While AI has the potential to facilitate progress, it can also lead to unethical use and weaken the integrity of academic research.

There are similar views in the literature on this subject. It is a fact that an academic who cannot control their emotions may resort to unethical behavior for achieving success (Maya, 2013).

However, due to the uniqueness of the subject of the study and the fact that it is a new field, opinions and suggestions regarding the relationship between the use of artificial intelligence and the emotional states of academics are not yet widely included in the literature. In this study, the claim that academics who cannot control their emotional state can achieve unfair success through the unfair use of artificial intelligence was also among the findings.

Within the framework of these findings, several concrete solution suggestions have been put forward. Firstly, continuous training sessions should be organized to enhance ethical awareness among academics and encourage personal ethical responsibility. Additionally, special ethical committees in the field of AI should be established, and academic publication and promotion committees should conduct more effective oversight. Furthermore, the development and implementation of specific regulations regarding the use of AI are crucial.

For future research, the involvement of ethical experts in evaluating the ethical consequences of AI use in academic research is essential. Moreover, psychology experts should conduct studies to better understand the relationship between academics' emotional states and professional advancement goals with the use of AI.

Further studies are needed to better understand the ethical implications of AI use in academic research. Ethical experts evaluating the ethical aspects of AI use and contributing to the improvement of regulations in this area are crucial. Additionally, through survey studies conducted by psychology experts with academics, it may be possible to better understand the impact of AI on emotional states and professional advancement goals.

In conclusion, an approach addressing the ethical issues of AI use in academia should be adopted. This indicates the need for increased ethical awareness in the academic community, improved institutional regulations, and more research. Addressing the ethical challenges of AI use in academia requires a multidisciplinary approach integrating ethical principles, psychological perspectives, and institutional regulations. In this way, the benefits of AI use can be maximized, while potential risks can be minimized.

9 Instead of conclusion: "Academic ethics, emotions and future?"

Since the term was first used in academic articles, in the field of AI, significant changes and transformations have been observed, allowing the direct use of AI by individuals in various fields. For example, AI usage is becoming increasingly prevalent in health applications, personal mobile phones, computers, cars, and many other areas and products (Roser, 2023). Like all fields, the use of AI in academia has become evident in recent years, leading to serious debates. While studies highlighting the positive aspects of using AI in academia exist, there are also studies indicating its negative effects (Bakiner, 2023).

This study specifically addresses the use of AI in academia, with a focus on investigating the relationship between academics' emotional states and unfair professional progression due to the use of AI.

The results of interviews conducted with ethics and psychology experts in this study lead to the following conclusions:

Ethics is based on the foundation of acting correctly. In this context, scientific ethics can be summarized as acting truthfully and honestly, not distorting data, and not trying to progress unfairly.

The use of AI in academia is becoming increasingly widespread. From a positive perspective, this usage significantly contributes to making studies more practical. However, it can lead to problems such as unfair authorship, devaluation of human authorship, and incorrect data.

The connection between academics' professional advancement goals and emotional states becomes prominent in this context. The potential of AI to facilitate progression can lead to unethical use.

To prevent such situations, it is recommended to organize training sessions to increase professional awareness, internalize ethics personally, establish ethical committees specific to the field of AI, conduct more effective audits by academic publication and promotion committees, and implement specific regulations for AI.

Finally, for future academic studies, it is suggested that the usage of AI in academic research be measured and evaluated by ethics experts. For psychologists, conducting surveys with academics to explore how they use AI in the context of their emotional states and professional advancement goals is recommended.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Ethics statement

The studies involving human participants were reviewed and approved by Near East University Scientific Research Ethics Committee. The participants provided written informed consent to participate in this study.

Author contributions

AD: Formal analysis, Methodology, Project administration, Supervision, Writing – review & editing. ACT: Data curation, Resources, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akcan, A. T., Malkoç, S., and Kızıltan, Ö. (2018). Akademisyenlere Göre Akademi ve Akademik Kültür (Academy and Academic Culture According to Faculty Members), *Bolu Abant İzzet Baysal University Journal of Faculty of Education*, 18, 569–591.

Altıntop, M. (2023). Academic text writing with artificial intelligence/smart learning technologies: the ChatGPT example. *J. Süleyman Demirel Univers. Institute of Soc. Sci.* 2, 186–211.

Arslan, K. (2020). Artificial Intelligence and Applicatıons in Education Western Anatolia. *Journal of Educational Sciences*, 11, 71–88.

Aydin, M. (2016). Etik Nedir? (What is Ethics?), Journal of Sakarya University Faculty of Theology, 18, 171–177.

Bakiner, O. (2023). What do academics say about artificial intelligence ethics? An overview of the scholarship. *AI and Ethics* 3, 513–525. doi: 10.1007/s43681-022-00182-4

Biernacki, P., and Waldorf, D. (1981). Snowball sampling: problems and techniques of chain referral sampling. *Sociol. Methods Res.* 10, 141–163. doi: 10.1177/004912418101000205

Chiu, T. K., Xia, Q., Zhou, X., Chai, C. S., and Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Comp. Educ.: Art. Intell.* 4:100118. doi: 10.1016/j.caeai.2022.100118

Crompton, H., and Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *Int. J. Educ. Technol. High. Educ.* 20, 1–22. doi: 10.1186/s41239-023-00392-8

Demir, E., Demir, C. G., and Özdemir, M. Ç. (2017). Akademik Yükseltme ve Atama Sürecine Yönelik Öğretim Üyesi Görüşleri (Faculty Members' Views on Academic Promotion and Appointment Process), *Journal of Higher Education and Science*, 1, 12–23.

Dolunay, A. (2018). *Dijital Çağda Yasal Ve Etik Kodlar Çerçevesinde Basın Hak Ve Özgürlükleri KKTC Örneği (freedom of the Press in the Digital age within the framework of legal and ethical codes: TRNC example.)* İstanbul: Oniki Levha Publishing.

Dolunay, A., and Kasap, F. (2018). Freedom of the press in the digital age within the frameworks of ethics, law and democracy education: example of the North Cyprus. *Qual. Quant.* 52, 663–683. doi: 10.1007/s11135-017-0645-x

Ercan, T., Daşlı, Y., and Biçer, B. (2021). Publishing ethics in scientific information. *CUJOSS* 45, 91–108.

Gasset, J. O. (2014). *Mission of the University*. New York: Routledge.

Girgin, A. (2000). *Yazılı Basında Haber ve Habercilik Etik'i (News and journalism ethics in the print media)*. İstanbul: İnkılâp Publishing.

Gürkan, T. (2018). Akademisyen Olmak (Being an academician), *Journal of Early Childhood Studies*, 2, 440–446.

Jabotinsky, H. Y., and Sarel, R. (2022). Co-authoring with an AI? Ethical dilemmas and artificial intelligence. *Ethical Dilemmas and Art. Intel., Arizona State Law J., Forthcoming*. 1–42. doi: 10.2139/ssrn.4303959

Johshon, J. M. (2002). "In-depth interviewing," in *Handbook of interview research context & method*. eds. J. F. Gubrium and J. A. Holstein. California: Sage Publications.

Kahn, R. L. (1983). *The dynamics of interviewing.* Florida: Robert E. Krieger Publishing Company.

Karataş, Z. (2017). Paradigm transformation in social sciences research: Rise of qualitative approach. *Turkish Journal of Social Work Research*, 1, 68–86.

Koçak, A., and Arun, Ö. (2013). The Sampling Problem in the Content Analysis Studies. *Journal of Selçuk Communication*, 4, 21–28. doi: 10.18094/si.51496

Kokina, J., and Davenport, T. H. (2017). The emergence of artificial intelligence: How automation is changing auditing, *Journal of emerging technologies in accounting*, 14, 115–122.

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., and Wang, Z. (2023). ChatGPT and a new academic reality: artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *J. Assoc. Inf. Sci. Technol.* 74, 570–581. doi: 10.1002/asi.24750

Maya, İ. (2013). Akademisyenlerin Meslek Ahlakına Aykırı Olan Davranışlara İlişkin Algıları: Çomü Eğitim Fakültesi Örneği (Academicians' Perceptions Of Behaviours Against Occupational Ethics: A Case In Comu, Faculty Of Education). *Turk. Stud.* 8, 491–509. doi: 10.7827/TurkishStudies.5039

McCracken, G. (1988). *The long interview*. California:Sage Publications.

Mijwil, M. M., Hiran, K. K., Doshi, R., Dadhich, M., Al-Mistarehi, A. H., and Bala, I. (2023). ChatGPT and the future of academic integrity in the artificial intelligence era: a new frontier. *Al-Salam J. Engineer. Technol.* 2, 116–127. doi: 10.55145/ajest.2023.02.02.015

Moressi, E. (2006). *News ethics establishment and criticism of moral journalism (Genç F.* Trans.). Ankara: Dost Publishing.

Muthukrishnan, N., Maleki, F., Ovens, K., Reinhold, C., Forghani, B., and Forghani, R. (2020). Brief history of artificial intelligence. *Neuroimaging Clinics* 30, 393–399. doi: 10.1016/j.nic.2020.07.004

Öztürk, K., and Şahin, M. E. (2018). Yapay sinir ağları ve yapay zekâ'ya genel bir bakış. (A General View of Artificial Neural Networks and Artificial Intelligence), *Takvim-i Vekayi*, 6, 25–36.

Păvăloaia, V. D., and Necula, S. C. (2023). Artificial intelligence as a disruptive technology—a systematic literature review. *Electronics* 12:1102. doi: 10.3390/electronics12051102

Peck, A.L, and Reel, G. (2013). *Media ethics at work, true stories from young professionals*. USA: SAGE Publications Ltd.

Punch, K. F. (2005). *Introduction to social research: Quantitative and qualitative approaches*. (D. Bayrak, H. B. Aslan and Z. Akyüz Trans.). Ankara: Siyasal Publishing.

Resnik, D. B. (2012). Ethical virtues in scientific research. *Account. Res.* 19, 329–343. doi: 10.1080/08989621.2012.728908

Roser, M. (2023). *The brief history of artificial intelligence: The world has changed fast– what might be next?. Our world in data*.

Sáiz-Manzanares, M. C., Almeida, L. S., Martín-Antón, L. J., Carbonero, M. A., and Valdivieso-Burón, J. A. (2022). Teacher training effectiveness in self-regulation in virtual environments. *Front. Psychol.* 13:776806. doi: 10.3389/fpsyg.2022.776806

Sariyasa, S., and Monika, K. A. L. (2023). Artificial intelligence and academic ethics in the era of Merdeka Belajar: how are Students' responses? *Jurnal Kependidikan: Jurnal Hasil Penelitian dan Kajian Kepustakaan di Bidang Pendidikan, Pengajaran dan Pembelajaran* 9, 986–995. doi: 10.33394/jk.v9i3.8720

Tekin, H. H. (2006). In-depth interview of qualitative research method as a data collection technique, *Istanbul University Journal of Sociology*, 3, 101–116.

Thunström, A. O. (2022). We asked GPT-3 to write an academic paper about itself—then we tried to get it published: an artificially intelligent first author presents many ethical questions—and could upend the publishing process. *Sci. Am.* 30. doi: 10.1038/scientificamerican0922-70

Tunç, B. (2007). Akademik unvan olgusu akademik yükseltme ve atama sürecinin değerlendirilmesi (the evaluation of the academic title, academic promotion and academic appointment). Unpublished Ph.D. Thesis, University of Ankara Education Sciences Institute, Ankara, Turkey.

Ülman, Y. I. (2006). "Bilimsel yayın etiği örneklerle bilimsel yanıltma türleri" in *Tıbbi yayın hazırlama kuralları ve yayın etiği*. eds. H. Yazıcı and M. Şenocak (İstanbul: Nobel Publishing), 49–61.

Uzun, R. (2007). İletişim Etiği Sorunlar ve Sorumluluklar (Communication ethics issues and responsibilities). *Ankara: Gazi University Library of Faculty of Communication*' 40. Year Publications.

Weber, P. R. (1990). *Basic Content Analysis*. Sage:London.

Yıldırım, A., and Şimşek, H. (2008). *Qualitative research methods in the social sciences*. Ankara: Seçkin Publishing.

Yördem, Y., and Şeker, H. (2018). Vıolatıons of publication Ethıcs and author right issues. *Dicle University Vocational School of Justice Dicle Justice J.* 2, 33–48.

frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 31 May 2024
DOI 10.3389/frai.2024.1398395

OPEN ACCESS

EDITED BY
Mariacarla Martí-González,
University of Valladolid, Spain

REVIEWED BY
Huma Shah,
Coventry University, United Kingdom
Jon Rueda,
University of Granada, Spain

*CORRESPONDENCE
Max Tretter
✉ max.tretter@fau.de

RECEIVED 09 March 2024
ACCEPTED 13 May 2024
PUBLISHED 31 May 2024

CITATION
Tretter M (2024) Equipping AI-decision-support-systems with emotional capabilities? Ethical perspectives. *Front. Artif. Intell.* 7:1398395. doi: 10.3389/frai.2024.1398395

Equipping AI-decision-support-systems with emotional capabilities? Ethical perspectives

Max Tretter*

Faculty of Humanities, Social Sciences, and Theology, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

It is important to accompany the research on Emotional Artificial Intelligence with ethical oversight. Previous publications on the ethics of Emotional Artificial Intelligence emphasize the importance of subjecting every (possible) type of Emotional Artificial Intelligence to separate ethical considerations. That's why, in this contribution I will focus on a particular subset of AI systems: AI-driven Decision-Support Systems (AI-DSS), and ask whether it would be advisable from an ethical perspective to equip these AI systems with emotional capacities. I will show, on one hand, equipping AI-DSS with emotional capabilities offers great opportunities, as they open the possibility to prevent emotionally biased decisions – but that it also amplifies the ethical challenges already posed by emotionally-incapable AI-DSS. Yet, if their introduction is accompanied by a broad social discourse and prepared by suitable measures to address these challenges, I argue, nothing should fundamentally stand in the way of equipping AI-DSS with emotional capabilities.

KEYWORDS
emotional intelligence, agency, responsibility, trust, emotion detection

1 Introduction

Emotional Artificial Intelligence (EAI) is a vibrant field of research (McStay, 2018; Misselhorn, 2021; Assunção et al., 2022). One of the main challenges in this field involves crafting AI-systems that are capable of analyzing human gestures, facial expressions, postures, speech, or behavior, and use this biometric data to accurately identify people's emotional states. This involves interpreting subtle biometric signals, including minor muscle movements or slight variations in vocal pitch, which may signal a range of emotions from stress and happiness to fear and sarcasm. Algorithms with the capability for such nuanced emotion detection are being researched across a range of settings, including *healthcare* (where EAI can be used to improve practitioner-patient interactions (Vagisha and Harendra, 2023) or mental health care (Joshi and Kanoongo, 2022)), *automotive safety* (where EAI is intended to detect signs of drowsiness or distraction and take safety measures (McStay and Urquhart, 2022)), or *education* (where attempts are being made to use EAI to improve pedagogical methods and respond better to the affective states of pupils (McStay, 2020a)). Further, in the realm of social robotics, there's ongoing research aimed at equipping robots with emotional capabilities, thereby enhancing their ability to engage empathetically and socially with humans (Marcos-Pablos and García-Peñalvo, 2022).

In light of the advancements achieved in the field of EAI over the recent years, along with the vast potential applications of emotionally-capable AI systems and the promising opportunities they offer, a number of studies have emerged offering the ethical perspectives on EAI (McStay, 2018, 2020b; Greene, 2020; Gremsl and Hödl, 2022; Ghotbi, 2023; Gossett, 2023). These investigations have highlighted the potential benefits of emotionally-capable AI systems, while also drawing attention to the associated risks, with key concerns including issues related to privacy, the potential for manipulation, and the threat of exacerbating socio-economic disparities. One central claim found in several ethical discussions on EAI is that the ethical evaluation of this technology hinges on its application context (e.g., healthcare, safety, or advertising) and its intended purpose (e.g., mitigating mental health issues, surveilling public areas, or boosting sales metrics) (Greene, 2020; Ghotbi, 2023).

Against this backdrop, I will focus on one specific type of AI systems: AI-driven Decision-Support-Systems (AI-DSS). These are algorithmic systems typically used in complex decision-making scenarios to analyze these situations with AI, including machine learning and predictive analytics, to deepen understanding, predict potential outcomes of various decision options, and offer data-driven recommendations to facilitate the decision-making process (Phillips-Wren, 2013). I will explore and ask whether it would be advisable from an ethical perspective to equip these AI-systems with emotional capacities. Despite the existence of a significant corpus of research that provides ethical perspectives on AI-DSS in general or their use in specific contexts (Braun et al., 2020; Lara and Deckers, 2020; Stefan and Carutasu, 2020; Cartolovni et al., 2022; Nikola et al., 2022), alongside a comprehensive body of literature addressing the ethics of EAI (McStay, 2018, 2020b; Greene, 2020; Gremsl and Hödl, 2022; Ghotbi, 2023; Gossett, 2023), so far, there has been no research that intersects these two domains. Specifically, there's a lack of investigation into the ethics of emotionally capable AI-DSS.

My goal is to bridge this gap and to argue that, on one hand, equipping AI-DSS with emotional capabilities offers great opportunities, as they open the possibility to prevent emotionally biased decisions, but that it also amplifies the ethical challenges already posed by emotionally-incapable AI-DSS. Yet, if their introduction is accompanied by a broad social discourse and prepared by suitable measures to address these challenges, I argue, nothing should fundamentally stand in the way of equipping AI-DSS with emotional capabilities.

To substantiate my thesis, I will first focus on the decision-making process, its complexities, and how AI-DSS can assist in making decisions. I will then examine the opportunities and risks associated with equipping these AI-DSS with emotional capabilities, discussing them, and making some suggestions about the advisability of emotionally-capable AI-DSS.

2 The difficulty of making decisions and the help of AI

Some decisions are easy to make. Others, however, are difficult. The level of difficulty often hinges on the number of people impacted and the potential severity of the outcomes. Decisions with minimal consequences that affect mainly oneself, such as choosing which pair of shoes to put on in the morning, tend to be simpler than life-altering choices like marriage, which involves other people and bears lasting repercussions. Furthermore, decision-making complexity also escalates with situational complexity and one's emotional state. A complex situation complicates the clarity of potential outcomes due to information scarcity, challenging the decision-making process (Dewey, 1929; Tretter, 2023). Emotional involvement further exacerbates this challenge, as too much emotion can skew perceptions and introduce biases (Mazzocco et al., 2019; Dorison et al., 2020).

The effects of strong emotional involvement on decision-making can be illustrated using an example from the military sector. Modern military operations are extremely complex and highly dynamic, requiring intricate coordination among various units like infantry, armor, artillery, air support, and logistics to ensure mutual support rather than interference. Furthermore, battlefield conditions can swiftly change, necessitating rapid responses to enemy maneuvers. This complexity and dynamics make strategic decision-making an extremely complicated matter – and can cause continuous emotional stress for the persons in charge. In situations where this stress intensifies, military personnel are more likely to misjudge situations, make a hasty decision, and thereby unduly endanger the lives of those affected (Gamble et al., 2018).

To assist decision-makers in such challenging situations, AI-DSS exist. Provided with sufficient high-quality data, such AI-DSS can quickly comprehend complex situations, analyze them, present possible options, and even simulate the outcomes of various decisions—thus offering recommendations on the most advisable course of action. Such systems are also available, e.g., for the military sector (Scharre, 2020; Szabadföldi, 2021), where AI-DSS are capable of assessing battlefield dynamics in fractions of a second, evaluating the level of threat, and recommending strategies tailored to specific situations. Through such advanced analysis and recommendation processes, AI-DSS significantly bolster the decision-making capacity of military personnel (Liao and Sun, 2020; Horyń et al., 2021).

3 The potential of emotionally capable AI-decision-support-systems

As just outlined, AI-DSS can assist in making complex decisions, taking into account a broad array of factors in their analysis, simulations, and advice. At present, however, they are limited by the fact that they cannot take into account the emotional disposition of decision-makers. This oversight is critical because, as demonstrated above, excessive emotional involvement can lead to misperceptions and misjudgments of situations, which in turn may result in hasty or biased decisions.

By equipping AI-DSS with emotional capabilities and enabling them to discern the emotional states of decision-makers, such as military personnel, which exceed the "normal" level of stress associated with such situations and tasks, this shortfall could be remedied. With the ability to assess users' emotional states, these AI systems could proactively alert individuals if their emotional engagement is likely to impair judgment, making them statistically more prone to errors and biased decisions. In situations where simple alerts might not suffice, the AI could recommend pausing decision-making processes until a more "balanced" emotional state is attained or suggest delegating their responsibilities temporarily. Equipping AI-DSS with emotional capabilities thus offers a forward-looking approach that promises to mitigate the risks of emotionally driven, biased decisions.

It is, no doubt, beneficial to detect and issue warnings about excessive emotional involvement. However, this should not mislead us into believing that emotions are inherently "negative" within the decision-making framework, or that decisions can or should be made on a purely rational basis (Seo and Barrett, 2007). In fact, while over-engagement of emotions can adversely affect decision-making, endeavors to entirely eliminate emotional influence from this process can be just as detrimental. As contemporary research in the field of emotions suggests, there's a symbiotic relationship between rational thought and emotions, debunking the notion that they are mutually exclusive (Damasio, 1994; Kappelhoff et al., 2019). Given this symbiotic relationship, attempts to exclude emotions from decision-making prove not only unrealistic but also disadvantageous for the decision-making process. This conclusion can be further underscored by everyday observations that, in certain scenarios, emotions can be favorable for decision-making (Mazzocco et al., 2019; Dorison et al., 2020; Gengler, 2020). For instance, worry or fear might prompt more thorough considerations in specific contexts, whereas empathy can lead to decisions that are more compassionate.

The ideal state for decision-making processes involves a "balanced" level of emotional engagement, where decision-makers strike a balance between being excessively emotionally involved and acting like emotionless robots. However, identifying what constitutes a "balanced" degree of emotional engagement in decision-making is complex, as the appropriate level of emotionality significantly varies by context and individual. Ideally, setting "thresholds" for emotional involvement should be personalized and contextual, presenting a substantial challenge. Until tailoring such specific thresholds becomes feasible, employing average benchmarks could serve as a practical interim strategy. This strategy could involve determining the typical degree of emotionality that different individuals demonstrate in specific situations (*situation-specific benchmarks*) or evaluating the general emotional responses of particular individuals across diverse scenarios (*individual-specific benchmarks*). While developing these benchmarks, AI-DSS can be just as useful as in checking, in specific decision-making scenarios, whether decision-makers are too emotionally involved (or not enough).

4 The challenges of emotionally capable AI-decision-support-systems

While endowing AI-DSS with emotional capabilities brings significant opportunities, it also raises complex challenges, beginning with the systems' functionality itself. Current emotionally-capable AI-systems often display biases related to culture, gender, age, and race. This predisposition allows for the precise detection of emotions in white, middle-aged men from Western backgrounds, whereas it fails to recognize with the same accuracy the emotions of individuals from diverse cultures, genders, ages, and racial backgrounds (Shimo, 2020; Kim et al., 2021; Ghotbi, 2023; Gossett, 2023). Yet, even in scenarios where emotionally-capable AI operates flawlessly, recognizing emotions across cultural, gender, age, and racial spectrums without bias, large challenges arise.

Notably, the challenges encountered with emotionally-capable AI-DSS mirror those associated with emotionally-incapable AI-DSS. I will argue that equipping AI-DSS with emotional capabilities exacerbates these existing challenges. In this context, I will particularly focus on the issue of *agency*, and then, building on this foundation, briefly explore the issues of *responsibility*, *accountability*, and *trust*.

One contentious topic in ethical discussions on AI-DSS is the issue of *agency* (Taddeo and Floridi, 2018; Jobin et al., 2019; Braun et al., 2020; Stefan and Carutasu, 2020; Cartolovni et al., 2022; Nikola et al., 2022). While AI-DSS are designed to *support* human decision-making through recommendations, leaving ultimate control with humans, the concern arises that AI's influence may subtly shift agency away from human decision-makers and toward the AI (Braun et al., 2020). For instance, consider a hypothetical scenario where a physician, despite their instinct or previous experience advocating for a different course of action, may be reticent to question a medical AI system's treatment suggestion. This reluctance could stem from the perception that the AI system is capable of analyzing a broader array of data, identifying more complex correlations, possessing a more current understanding of medical literature, and executing thorough simulations (Tretter, 2023). This scenario, and similar examples could be found for other contexts, illustrates how difficult it can become for people to contradict the recommendations of AI-DSS and that it may be the easier path to simply agree with AI recommendations. This trend, however, if left unchecked, could gradually erode human agency within the decision-making process.

The hurdles to challenging AI-DSS intensify significantly when individuals, upon deciding against an AI's recommendation, are subsequently required to justify their decision. In such cases, relying on personal intuition or past experiences may not be considered adequate justification. Confronted with these daunting barriers to overlooking AI suggestions, individuals may increasingly find themselves in a position where they merely validate and approve the proposals of AI-DSS, marking a significant shift in decision-making agency toward AI (Tretter, 2023; Tretter et al., 2023).

Therefore, it is evident that AI-DSS, even without emotional capabilities, can significantly impact user decisions and gradually encroach upon decision-making agency. Incorporating emotional capabilities into AI-DSS may further amplify this issue. Where individuals find themselves having to justify decisions that deviate from AI-generated advice, they now encounter the additional risk that their divergent choices might be attributed to their emotional state. This could further deter people from questioning and deciding against AI recommendations, deepening concerns over the erosion of agency.[1]

Where agency is increasingly challenged by emotionally-capable AI-DSS, this has far-reaching consequences for other issues. If the agency in a decision clearly lies with the human, they can be morally responsible for those decisions and legally liable for their outcomes. However, as humans relinquish more agency, for example, because AI

[1] Another consideration is that emotionally-capable AI-DSS could tailor their recommendations precisely to the user's emotional state, allowing them to nudge users toward specific decisions with unparalleled accuracy. The practice of nudging, due to its highly manipulative nature (Sunstein, 2015), remains ethically questionable whether it is carried out by AI-DSS with or without emotional capabilities (Fritzen, 2023). However, the capacity for such nudging is significantly enhanced when employed by emotionally-capable AI-DSS, intensifying the challenge concerning agency.

systems significantly influence or even manipulate their decisions or make decisions independently, the less they can legitimately be held responsible and liable. This raises the crucial question of where responsibility and liability should then lie: with the developers of AI-DSS, the institutions that deploy or individuals that use them, the AI system itself, all of them together, or none at all? While such issues of responsibility and liability have been extensively debated in contexts like self-driving cars (Coeckelbergh, 2016; Gless et al., 2016), smart healthcare (Smith, 2021; Sand et al., 2022), and autonomous weapons systems (Santoni de Sio and van den Hoven, 2018; Wood, 2023), no satisfactory resolution has yet emerged. And it is expected that this discussion will become even more complex when AI-DSS are equipped with emotional capacities.

Where responsibility and liability are increasingly called into question by AI-DSS equipped with emotional capabilities, the question arises about the impact this has on existing trust toward these systems. Will trust increase because they can now account for emotional aspects, enabling more thoughtful and sensitive support? Will trust in them decrease due to the heightened risk of unnoticed manipulation by their emotional capabilities? Or will these enhancements have no effect on trust? Further, given that these systems operate within complex sociotechnical frameworks (Schmidl, 2022), the question also arises as to how shifts in trust toward AI-DSS will influence trust toward the domains and institutions deploying them (Samhammer et al., 2023; Tretter et al., 2023).

These concerns about responsibility, liability, and trust are, as hinted above, already relevant in the context on emotionally-incapable AI-DSS. Nevertheless, the extent to which AI systems encroach upon human agency – significantly more so in the case of emotionally-capable AI-DSS than their emotionally-incapable counterparts – amplifies the scrutiny on these follow-up issues. That's why emotionally-capable AI-DSS intensify these concerns about responsibility, liability, and trust even more.

5 Discussion

Considering the opportunities that emerge, alongside the heightened challenges of equipping AI-DSS with emotional capabilities, the question of whether emotionally-capable AI-DSS are ethically advisable cannot be simply answered with a straightforward "yes" or an unequivocal "no." On one side, dismissing the potential benefits of providing AI-DSS with emotional capabilities by outright rejecting the concept of emotionally-capable AI-DSS would be negligent. Such a choice would ignore the opportunity to mitigate emotionally biased judgments and decisions, potentially risking lives in critical situations (e.g., in the military context).

On the other side, it would be equally negligent to overlook the risks involved and to unconditionally support equipping AI-DSS with emotional capabilities. Opting for this path would fail to address the peril of agency progressively shifting from humans to AI, exacerbating subsequent responsibility gaps, lack of liability, and serious trust issues.

From an ethical perspective, the question of whether AI-DSS should be equipped with emotional capacities might best be answered with a "yes, but…." When a broad societal debate is conducted, in which all perspectives are welcome to deliberate the contexts and manners in which emotionally-capable AI-DSS should be utilized, and if precautionary measures are established from the outset to prevent the loss of human agency, responsibility, and trust, nothing is fundamentally standing in the way of equipping AI-DSS with emotional capabilities. However, this approval remains valid only so long as these stipulations are genuinely fulfilled. Failing to meet these criteria transforms the "yes, but…" into a "no, unless…." As is often the case, the devil lies in the details of execution.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

MT: Writing – review & editing, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study received funding from the Federal Ministry of Education and Research (Bundesministeriums für Forschung und Bildung, Grant Number: 01GP2202B) and the German Research Foundation (Deutschen Forschungsgemeinschaft, SFB 1483 – Project-ID 442419336). The funders had no involvement in the research process or in the drafting of this manuscript.

Conflict of interest

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Assunção, G., Patrão, B., Castelo-Branco, M., and Menezes, P. (2022). An overview of emotion in artificial intelligence. *IEEE Trans. Artif. Intell.* 3, 867–886. doi: 10.1109/TAI.2022.3159614

Braun, M., Hummel, P., Beck, S., and Dabrock, P. (2020). Primer on an ethics of AI-based decision support systems in the clinic. *J. Med. Ethics* 47:e3. doi: 10.1136/medethics-2019-105860

Cartolovni, A., Tomicic, A., and Lazic Mosler, E. (2022). Ethical, legal, and social considerations of AI-based medical decision-support tools: a scoping review. *Int. J. Med. Inform.* 161:104738. doi: 10.1016/j.ijmedinf.2022.104738

Coeckelbergh, M. (2016). Responsibility and the moral phenomenology of using self-driving cars. *Appl. Artif. Intell.* 30, 748–757. doi: 10.1080/08839514.2016.1229759

Damasio, A. (1994). *Descartes' error. Emotion, reason and the human brain*. London: Random House.

Dewey, J. (1929). *The quest for certainty. A study of the relation of knowledge and action*. London: George Allen & Unwin.

Dorison, C. A., Klusowski, J., Han, S., and Lerner, J. S. (2020). Emotion in organizational judgment and decision making. *Organ. Dyn.* 49:100702. doi: 10.1016/j.orgdyn.2019.02.004

Fritzen, N. M. (2023). *AI-nudging and individual autonomy: Moral permissibility and policy recommendations*. Vienna: Central European University.

Gamble, K. R., Vettel, J. M., Patton, D. J., Eddy, M. D., Caroline Davis, F., Garcia, J. O., et al. (2018). Different profiles of decision making and physiology under varying levels of stress in trained military personnel. *Int. J. Psychophysiol.* 131, 73–80. doi: 10.1016/j.ijpsycho.2018.03.017

Gengler, A. M. (2020). Emotions and medical decision-making. *Soc. Psychol. Q.* 83, 174–194. doi: 10.1177/0190272519876937

Ghotbi, N. (2023). The ethics of emotional artificial intelligence: a mixed method analysis. *Asian Bioeth. Rev.* 15, 417–430. doi: 10.1007/s41649-022-00237-y

Gless, S., Silverman, E., and Weigend, T. (2016). If robots cause harm, who is to blame? Self-driving cars and criminal liability. *New Crim. L. Rev.* 19, 412–436. doi: 10.1525/nclr.2016.19.3.412

Gossett, S. (2023). Emotion AI: 3 experts on the possibilities and risks. Available at: https://builtin.com/artificial-intelligence/emotion-ai

Greene, G.. (2020). The ethics of AI and emotional intelligence. Available at: https://partnershiponai.org/paper/the-ethics-of-ai-and-emotional-intelligence/

Gremsl, T., and Hödl, E. (2022). Emotional AI: legal and ethical challenges. *Inf. Polity* 27, 163–174. doi: 10.3233/IP-211529

Horyń, W., Bielewicz, M., and Joks, A. (2021). "AI-supported decision-making process in multidomain military operations" in *Artificial intelligence and its contexts: security, business and governance*. eds. A. Visvizi and M. Bodziany (Cham: Springer International Publishing), 93–107.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 1, 389–399. doi: 10.1038/s42256-019-0088-2

Joshi, M. L., and Kanoongo, N. (2022). Depression detection using emotional artificial intelligence and machine learning: a closer review. *Mater. Today Proc.* 58, 217–226. doi: 10.1016/j.matpr.2022.01.467

Kappelhoff, H., Bakels, J.-H., Lehmann, H., and Schmitt, C. (2019). *Emotionen. Ein interdisziplinäres Handbuch*. Stuttgart: J.B. Metzler.

Kim, E., Bryant, D. A., Srikanth, D., and Howard, A.. (2021). Age Bias in emotion detection: an analysis of facial emotion recognition performance on young, middle-aged, and older adults. In Paper presented at the Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery. New York

Lara, F., and Deckers, J. (2020). Artificial intelligence as a Socratic assistant for moral enhancement. *Neuroethics* 13, 275–287. doi: 10.1007/s12152-019-09401-y

Liao, X., and Sun, Z. H.. (2020) Research on combat deduction platform technology for intelligent operational decision. in *Proceedings of 2019 Chinese Intelligent Automation Conference*, ed. Z. Dog (Singapore: Springer), 1–13.

Marcos-Pablos, S., and García-Peñalvo, F. J. (2022). *Emotional intelligence in robotics: a scoping review*. Cham: Springer.

Mazzocco, K., Masiero, M., Carriero, M. C., and Pravettoni, G. (2019). The role of emotions in cancer patients' decision-making. *Ecancermedicalscience* 13:914. doi: 10.3332/ecancer.2019.914

McStay, A. (2018). *Emotional AI: the rise of empathic media*. London, Thousand Oaks: Sage.

McStay, A. (2020a). Emotional AI and EdTech: serving the public good? *Learn. Media Technol.* 45, 270–283. doi: 10.1080/17439884.2020.1686016

McStay, A. (2020b). Emotional AI, soft biometrics and the surveillance of emotional life: an unusual consensus on privacy. *Big Data Soc.* 7:205395172090438. doi: 10.1177/2053951720904386

McStay, A., and Urquhart, L. (2022). In cars (are we really safest of all?): interior sensing and emotional opacity. *Int. Rev. Law Comput. Technol.* 36, 470–493. doi: 10.1080/13600869.2021.2009181

Misselhorn, C. (2021). *Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & co*. Ditzingen: Reclam Verlag.

Nikola, B.-A., Andrea, F., Susanne, J., Tanja, K., Federico, M., Phyllis, B., et al. (2022). AI support for ethical decision-making around resuscitation: proceed with care. *J. Med. Ethics* 48, 175–183. doi: 10.1136/medethics-2020-106786

Phillips-Wren, G. (2013). "Intelligent decision support systems" in *Multicriteria decision aid and artificial intelligence: Links, theory and applications*. eds. M. Doumpos and E. Grigoroudis (Chichester: Wiley), 25–44.

Samhammer, D., Beck, S., Budde, K., Burchardt, A., Faber, M., Gerndt, S., et al. (2023). *Klinische Entscheidungsfindung mit Künstlicher Intelligenz. Ein interdisziplinärer Governance-Ansatz*. Berlin, Heidelberg: Springer.

Sand, M., Durán, J. M., and Jongsma, K. R. (2022). Responsibility beyond design: physicians' requirements for ethical medical AI. *Bioethics* 36, 162–169. doi: 10.1111/bioe.12887

Santoni de Sio, F., and van den Hoven, J. (2018). Meaningful human control over autonomous systems: a philosophical account. *Front. Robot. AI* 5:15. doi: 10.3389/frobt.2018.00015

Scharre, P. (2020). *Army of none: autonomous weapons and the future of warfare*. New York, London: W. W. Norton & Company.

Schmidl, A. (2022). *Relationen. Eine postphänomenologische Soziologie der Körper, Technologien und Wirklichkeiten*. Weilerswist: Velbrück Wissenschaft.

Seo, M. G., and Barrett, L. F. (2007). Being emotional during decision making-good or bad? *Acad. Manage. J.* 50, 923–940. doi: 10.5465/amj.2007.26279217

Shimo, S. (2020), Risks of bias in AI-based emotional analysis technology from diversity perspectives. In 2020 IEEE International Symposium on Technology and Society (ISTAS). Tempe, AZ, USA: IEEE

Smith, H. (2021). Clinical AI: opacity, accountability, responsibility and liability. *AI & Soc.* 36, 535–545. doi: 10.1007/s00146-020-01019-6

Stefan, R., and Carutasu, G. (2020). "How to approach ethics in intelligent decision support systems" in *Innovation in sustainable management and entrepreneurship*. eds. G. Prostean, J. Lavios Villahoz, L. Brancu and G. Bakacsi (Cham: Springer), 25–40.

Sunstein, C. R. (2015). The ethics of nudging. *Yale J. Regul.* 32, 413–450,

Szabadföldi, I. (2021). Artificial intelligence in military application–opportunities and challenges. *Land Forces Acad. Rev.* 26, 157–165. doi: 10.2478/raft-2021-0022

Taddeo, M., and Floridi, L. (2018). How AI can be a force for good. *Science* 361, 751–752. doi: 10.1126/science.aat5991

Tretter, M. (2023). "Ambivalenzen gegenwärtiger Gewissheitsbestrebungen. Menschliche Entscheidungsfreiheit in einer gewisserwerdenden Welt" in *Alexa, wie hast du's mit der Religion? Interreligiöse Zugänge zu Technik und Künstlicher Intelligenz*. eds. A. Puzio and N. Kunkel (Darmstadt: wbg – Wissen. Bildung. Gemeinschaft), 135–156.

Tretter, M., Ott, T., and Dabrock, P. (2023). AI-produced certainties in health care: current and future challenges. *AI Ethics* 4:6. doi: 10.1007/s43681-023-00374-6

Tretter, M., Samhammer, D., and Dabrock, P. (2023). Künstliche Intelligenz in der Medizin: Von Entlastungen und neuen Anforderungen im ärztlichen Handeln. *Ethik Med.* 36, 7–29. doi: 10.1007/s00481-023-00789-z

Vagisha, S., and Harendra, K. (2023). Emotional intelligence in the era of artificial intelligence for medical professionals. *Int. J. Med. Grad.* 2:112. doi: 10.56570/jimgs.v2i2.112

Wood, N. G. (2023). Autonomous weapon systems and responsibility gaps: a taxonomy. *Ethics Inf. Technol.* 25:16. doi: 10.1007/s10676-023-09690-1

frontiers | Frontiers in Psychology

TYPE Systematic Review
PUBLISHED 19 June 2024
DOI 10.3389/fpsyg.2024.1387089

OPEN ACCESS

EDITED BY
Maria Clelia Zurlo,
University of Naples Federico II, Italy

REVIEWED BY
Michela Ponticorvo,
University of Naples Federico II, Italy
Enrique H. Riquelme,
Temuco Catholic University, Chile

*CORRESPONDENCE
Mariacarla Martí-González
✉ mariacarla.marti@gmail.com

RECEIVED 16 February 2024
ACCEPTED 22 May 2024
PUBLISHED 19 June 2024

CITATION
Vistorte AOR, Deroncele-Acosta A, Ayala JLM, Barrasa A, López-Granero C and Martí-González M (2024) Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review. *Front. Psychol.* 15:1387089.
doi: 10.3389/fpsyg.2024.1387089

Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review

Angel Olider Rojas Vistorte[1,2], Angel Deroncele-Acosta[3], Juan Luis Martín Ayala[1,2], Angel Barrasa[4], Caridad López-Granero[4] and Mariacarla Martí-González[5]*

[1]Psychology Department, European University of the Atlantic, Santander, Spain, [2]Psychology Department, International Ibero-American University, Mexico, Mexico, [3]Escuela de Postgrado, Universidad San Ignacio de Loyola, Lima, Peru, [4]Department of Psychology and Sociology, University of Zaragoza, Teruel, Spain, [5]Department of Social Anthropology and Social Psychology, Complutense University of Madrid, Madrid, Spain

Introduction: Artificial Intelligence (AI) is transforming multiple sectors within our society, including education. In this context, emotions play a fundamental role in the teaching-learning process given that they influence academic performance, motivation, information retention, and student well-being. Thus, the integration of AI in emotional assessment within educational environments offers several advantages that can transform how we understand and address the socio-emotional development of students. However, there remains a lack of comprehensive approach that systematizes advancements, challenges, and opportunities in this field.

Aim: This systematic literature review aims to explore how artificial intelligence (AI) is used to evaluate emotions within educational settings. We provide a comprehensive overview of the current state of research, focusing on advancements, challenges, and opportunities in the domain of AI-driven emotional assessment within educational settings.

Method: The review involved a search across the following academic databases: Pubmed, Web of Science, PsycINFO and Scopus. Forty-one articles were selected that meet the established inclusion criteria. These articles were analyzed to extract key insights related to the integration of AI and emotional assessment within educational environments.

Results: The findings reveal a variety of AI-driven approaches that were developed to capture and analyze students' emotional states during learning activities. The findings are summarized in four fundamental topics: (1) emotion recognition in education, (2) technology integration and learning outcomes, (3) special education and assistive technology, (4) affective computing. Among the key AI techniques employed are machine learning and facial recognition, which are used to assess emotions. These approaches demonstrate promising potential in enhancing pedagogical strategies and creating adaptive learning environments that cater to individual emotional needs. The review identified emerging factors that, while important, require further investigation to understand their relationships and implications fully. These elements could significantly enhance the use of AI in assessing emotions within educational settings. Specifically, we are referring to: (1) federated learning, (2) convolutional neural network (CNN), (3) recurrent neural network (RNN), (4) facial expression databases, and (5) ethics in the development of intelligent systems.

Conclusion: This systematic literature review showcases the significance of AI in revolutionizing educational practices through emotion assessment. While advancements are evident, challenges related to accuracy, privacy, and cross-cultural validity were also identified. The synthesis of existing research highlights the need for further research into refining AI models for emotion recognition and emphasizes the importance of ethical considerations in implementing AI technologies within educational contexts.

KEYWORDS

emotions, artificial intelligence, teaching-learning, education, assessment

1 Introduction

The integration of Artificial Intelligence (AI) into educational settings marks a significant advancement in detecting, assessing, and nurturing students' emotions. AI's ability to analyze complex emotional behavior patterns through data collected during the learning process enables a deeper understanding of each student's needs.

By employing advanced algorithms, AI can detect signs of frustration, boredom, or enthusiasm, allowing educators to tailor their teaching methods more effectively. Additionally, AI can provide instant, personalized feedback based on emotional analysis, thereby creating a learning environment that is more attuned to students' emotional well-being. This comprehensive approach significantly contributes to students' holistic development, enhancing their ability to manage emotions, build positive relationships, and improve their academic performance.

In this regard D'Mello and Graesser (2012) raise that AI can predict student emotions (boredom, fluency/engagement, confusion, and frustration) by analyzing the text of dialogues between students and tutors during interactions with an "Intelligent Tutoring System." These AI-driven intelligent tutoring systems can positively influence student motivation by incorporating artificially intelligent educational models, such as the "Mobile Adaptive Personalized Learning Environment" -MAPLE- (Mehigan and Pitt, 2019). Thus, artificial tutors with synthesized emotions can adapt their behavior to students' reactions and affective states, improving their performance in e-learning systems (Florea and Kalisz, 2005).

Another interesting study argues that AI can help detect and assess students' emotions within interactive digital learning environments (IDLE) and adapt the environment accordingly to meet their real needs, potentially improving learning (Arguel et al., 2019). AI may also classify students' emotions during their interaction with immersive environments, allowing for a better understanding of their emotional experiences (Rodríguez et al., 2020).

AI can also analyze emotions from text, enhancing student motivation and performance in e-learning environments (Rodriguez et al., 2012). Simultaneously, it can gauge the intensity of emotions and tailor lessons to individual needs, promoting successful completion of academic studies (Sumithra et al., 2022). Similarly, a recent study found that using deep learning methods to detect students' emotions can significantly boost productivity and enhance the educational process (AlZu'bi et al., 2022).

In a systematic review, de Oliveira and Rodrigues (2021) discovered that 60% of recent studies on human behavior and AI, specifically from the past three and a half years, focus on emotion-driven organizations. This trend highlights the growing interest and novelty of the field.

Among the efforts to incorporate AI into emotional management within educational settings, the "Biologically Inspired Cognitive Architecture" (eBICA) is notable. Developed by Samsonovich (2020), eBICA allows AI to understand and interact with human emotions during social interactions. Additionally, the emotion-based artificial decision-making model has been shown to enhance the performance of educational agents in virtual settings (Yang and Zhen, 2014). Another approach involves the integration of emotional agents in AI-based learning environments to improve learner motivation, self-assessment, and self-motivation by improving the socioemotional climate (Gorga and Schneider, 2009), especially affective computing (Kort et al., 2001; González-Hernández et al., 2018; Ninaus et al., 2019; Shobana and Kumar, 2021; He et al., 2022; Aly et al., 2023; Villegas-Ch et al., 2023).

Recent advancements reveal that artificial intelligence (AI) can not only recognize but also predict emotions (Alm et al., 2005; Lin et al., 2023; Singh et al., 2023). This capability extends beyond identifying current emotional states, enabling systems such as virtual assistants and Intelligent Tutoring Systems (ITS) to proactively adapt and respond more effectively to students' emotional needs, thus enhancing the learning experience.

AI also significantly impacts social emotions such as empathy, compassion, and interpersonal phenomena like justice and cooperation, which are crucial for learning (Lamm and Singer, 2010).

Furthermore, AI can analyze empathic behavior in dynamic social contexts like educational settings. There are now models that use deep learning to foster emotional intelligence, processing multimodal emotional signals to generate appropriate empathic responses (Alanazi et al., 2023).

Overall, despite the challenges associated with AI's empathic abilities, it is acknowledged that AI offers valuable tools for promoting empathic skills, essential for social cooperation, and ethical and prosocial behavior (Gómez-León, 2022).

The importance of AI in supporting mental health is finally recognized, an area supported by hundreds of progressively increasing studies (Mohr et al., 2017; Garcia-Ceja et al., 2018; Graham et al., 2019; Shatte et al., 2019) taking into account that AI systems can provide emotional support and personalized advice to students and other educational actors experiencing stress or depression and provide advice and feedback based on emotional well-being.

1.1 Intelligent tutoring systems and emotions

Intelligent Tutoring Systems (ITS) are closely related to students' emotions, since learning and emotions are an inseparable binomial. This is expressed in the cognitive-affective unity of the human personality. Intelligent tutoring systems are evolving to address not only the cognitive aspect of learning, but also the emotional needs of students to improve their educational experience and performance. In this sense, configurations are being incorporated that enable ITS to detect emotions, content adaptation, emotional support, and personalized feedback, moving toward an emotionally intelligent tutoring system (Mohanan et al., 2018).

A study involving "MetaTutor," a hypermedia-based intelligent tutoring system (ITS), showcases the capabilities of ITS to enhance learning experiences. MetaTutor provides students with feedback on the impact of positive and negative emotions during learning. It also guides students on how to regulate specific emotions to optimize learning effectiveness. Importantly, MetaTutor assesses not only cognitive processes but also metacognitive processes, emphasizing its comprehensive approach to student learning and emotional management (Taub et al., 2021).

A review study on emotion regulation in intelligent tutoring systems (ITS) highlights a consensus among researchers in computerized learning. It suggests that ITS could greatly enhance their effectiveness if they were able to adapt to the emotional states of students. This adaptation would allow ITS to better support personalized learning experiences by responding dynamically to the emotional and cognitive needs of each student (Malekzadeh et al., 2015).

There is a growing body of research linking intelligent tutoring systems (ITS) to emotion during the learning process. Among the most significant advances is the analysis of facial expressions to estimate the emotional state of a student using ITS (Sarrafzadeh et al., 2003); the relationship between emotion variability, self-regulated learning and task performance in ITS (Li S. et al., 2021; Li W.-C. et al., 2021); inducing positive emotional states in ITS (Chaffar et al., 2009); a new approach toward model students' socio-emotional attributes to predict their performance in ITS (Assielou et al., 2021); the integration of emotion management strategies in ITS (Malekzadeh et al., 2014); emotional pedagogical agents in ITS (Sun et al., 2013); the use of emotional coping strategies in ITS (Chaffar and Frasson, 2010); among many other results that clearly show that ITS have a close link with human emotions.

2 Methods and procedures

We performed a systematic review of the scientific literature through the following databases: Pubmed, Web of Science, PsycINFO and Scopus. These articles were analyzed to extract key insights related to the integration of AI and emotional assessment within educational environments. Additionally, reference lists of included studies and reviews were checked for potentially relevant articles not identified through the electronic search.

The identification of thematic clusters was carried out through a process of analysis and synthesis of the studies included in the review. The criteria used were the following:

1. *Thematic frequency*: This criterion allowed us to identify the frequency with which certain themes or concepts appeared in the studies reviewed. This involved searching for and recording patterns of key terms in the titles, abstracts, keywords and sections of the studies reviewed.
2. *AI technology used*: This criterion is based on the specific artificial intelligence technologies used in the studies reviewed. It involves a detailed analysis of the techniques, tools, algorithms and technological approaches used for the evaluation of emotions in educational environments.
3. *Domain or scope of application*: This criterion focused on the specific contexts in which artificial intelligence technologies were applied to evaluate emotions in educational settings, including special education. It examines whether the studies focused on particular areas such as general education, vocational training or distance learning, as well as special education for students with special educational needs.
4. *Results*: This criterion allowed us to examine the findings of each study, especially in relation to the relevant aspects for the integration of artificial intelligence in the evaluation of emotions in educational environments. The observed effects, conclusions reached and implications for educational practice were considered.

Once these criteria have been determined, we continue with the process of identifying thematic clusters, following the following seven-step procedure:

1. *Study selection*: We began with an exhaustive search of the relevant literature using academic databases and specialized search engines. Predefined inclusion and exclusion criteria were applied to select relevant studies that addressed the topic of integrating artificial intelligence to assess emotions in educational settings.
2. *Information extraction*: Based on the established criteria, the research team began the process of extracting key information from each selected study, such as recurring concepts and processes; Applied AI technology, intervention context, results and main conclusions. This information provided a solid basis for analysis and comparison between studies.
3. *Identification of emerging themes*: All the extracted information was examined to identify recurring themes and organize emerging patterns related to the integration of artificial intelligence and the evaluation of emotions in educational environments. This involved a rereading of each study with the extracted information to understand its content and context.
4. *Data Coding*: Codes or labels were assigned to each emerging theme or pattern.
5. *Grouping into thematic clusters*: Using the codes assigned to each study or fragment, the codes were grouped into coherent thematic clusters. This process involved identifying similarities and relationships between the coded information and organizing them into groups that address specific aspects of integrating artificial intelligence to assess emotions in educational settings.
6. *Refinement and validation*: Thematic clusters were reviewed and refined to ensure consistency and relevance. Cross-validation was carried out between the researchers involved in the review to ensure accuracy and consistency in the grouping of studies. At first, 9 clusters had been formed, however, this

process allowed for greater integration, managing to refine and achieve 4 thematic clusters.

7 *Analysis and synthesis*: Once thematic clustering was completed, a detailed analysis of the studies within each cluster was conducted to identify trends, discrepancies, and notable areas of interest. This stage allowed us to synthesize the information collected and provide a contextualized view of the literature reviewed on the topic.

These thematic clusters were organized with the objective of providing a coherent structure to analyze and synthesize the information collected, thus facilitating the understanding of trends and advances in the integration of artificial intelligence to evaluate emotions in educational environments.

We used the following search terms: artificial intelligence terms AND recognition of emotions AND educational context terms as follows (Figure 1):

Cluster 1: "artificial intelligence" OR "machine intelligence" OR "intelligent support" OR "intelligent virtual reality" OR "chat bot*" OR "machine learning" OR "automated tutor" OR "personal tutor*" OR "intelligent agent*" OR "expert system" OR "neural network" OR "natural language processing."

Cluster 2: "Emotion recognition" OR "Speech Emotion Recognition" OR "Emotion Classification" OR "Emotional State" OR "Facial Emotion Recognition" OR "Facial Emotions" OR "Emotion Detection" OR "Emotionality" OR "Human Emotion" OR "Emotional Speech" OR "Multimodal Emotion Recognition" OR "Emotional Intelligence" OR "Automatic Emotion Recognition" OR "Human Emotion Recognition" OR "Emotion Analysis."

Cluster 3: "educational" OR "educational environments" OR "learning environments" OR "Educational Settings" OR "educational context" OR "pedagogical environments" OR "academic settings" OR "classroom environments" OR "learning spaces" OR "educational institutions" OR "school environments" OR "educational facilities" OR "teaching and learning environments" OR "educational institutions" OR "school systems" OR "academic programs" OR "higher education" OR "pedagogical approaches" OR "university campuses."

Titles and abstracts were screened, and full reports of potentially relevant studies were obtained using a Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) model (Page et al., 2021). Two authors (AORV and AD) independently assessed the reports for eligibility, with discrepancies resolved by discussion with a third author (JL).

We included quantitative studies in English, Spanish, and Portuguese, and studies related to both virtual and face-to-face educational environments. Articles were excluded based on the following exclusion criteria: (1) if they referred to non-data-based studies (e.g., editorials, commentaries, opinion papers, and review papers), and (2) if stigmatizing attitudes were assessed among non-physician primary care professionals, such as nurses, technicians, social workers, and other professionals, among mental health professionals, or among the general population. Data on study design, sample characteristics, and findings were extracted independently by three authors (MC, CAF and JLMA). Because of the heterogeneity between studies, which hindered a statistical synthesis of their results, we summarized evidence from articles included in the review through a narrative synthesis (Popay et al., 2006).

3 Results and discussion

One thousand and fifteen articles were identified in the four databases (Scopus: 366, Web of Science: 203, Pubmed: 163 and PsycINFO: 283). 135 articles were identified as potentially relevant and were assessed against eligibility criteria. Forty-one studies fulfilled

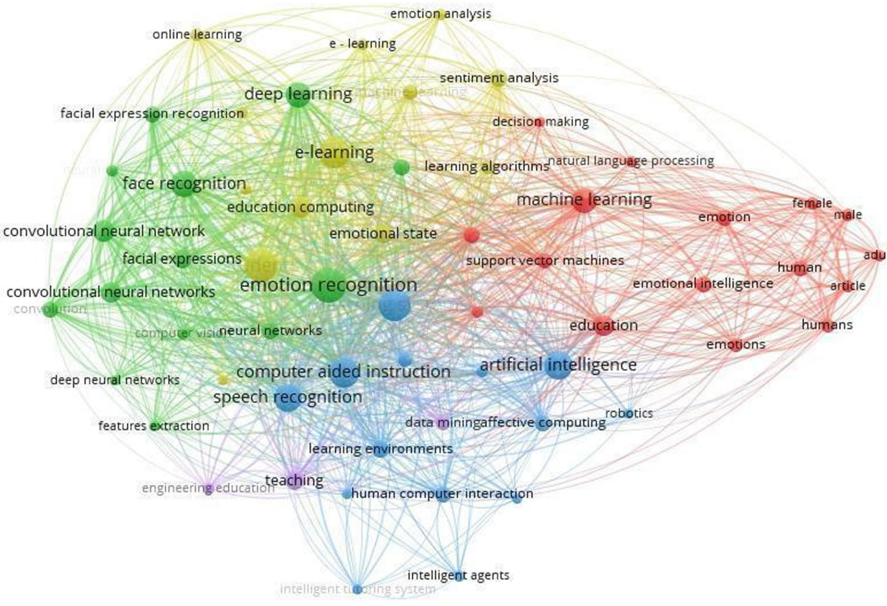


FIGURE 1
Co-occurrence network. VOSviewer_1.6.18.

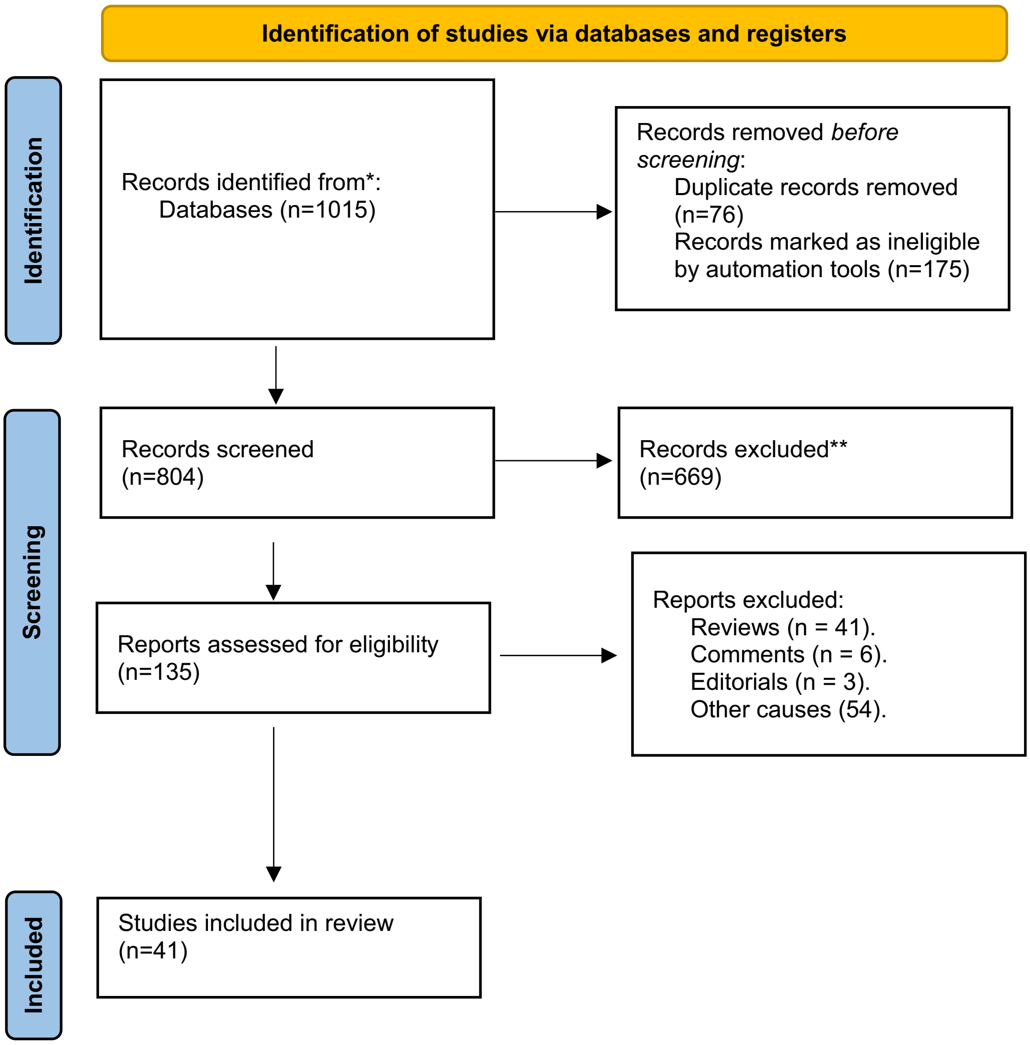

Identification of studies via databases and registers
Identification
Screening
Included
Records identified from*:
Databases (n=1015)
Records removed *before screening*:
Duplicate records removed (n=76)
Records marked as ineligible by automation tools (n=175)
Records screened (n=804)
Records excluded** (n=669)
Reports assessed for eligibility (n=135)
Reports excluded:
Reviews (n = 41).
Comments (n = 6).
Editorials (n = 3).
Other causes (54).
Studies included in review (n=41)

FIGURE 2
The number of articles found to be included in the review (Page et al., 2021).

inclusion criteria using the PRISMA model (Figure 2) and are summarized in Supplementary Table S1.

Every study included in the review used a cross-sectional design or used databases to investigate the integration of artificial intelligence (AI) for emotional assessment within educational contexts. Nine studies were conducted in China, one in Colombia, Ecuador, Egypt, Germany, seven in India, two in Iran, one in Japan and Jordan, two in Morocco, two in Russia and Spain, one in Thailand, Tunisia and United Arab Emirates, and five in the United States.

Based on the results of the articles included in the review, we can consider that there are several points in common and others that are more specific. We present the results and discussion based on each of these topics: *emotion recognition in education, technology integration and learning outcomes, special education and assistive technology, affective computing.*

3.1 Emotion recognition in education

Emotion recognition is essential for understanding how emotions affect peer interactions, academic performance, and engagement in online and virtual learning environments (Standen et al., 2020; Dehbozorgi and Kunuku, 2023; Villegas-Ch et al., 2023).

The main methods used in the research to analyze students' emotional states were related to facial expressions, eye movements, and biosignal data (Ninaus et al., 2019; Dehbozorgi and Kunuku, 2023; Villegas-Ch et al., 2023; Yugal et al., 2023). During online lessons, monitoring systems studied real-time attention, emotions and feelings (He et al., 2022; Dehbozorgi and Kunuku, 2023).

In educational settings, the use of artificial intelligence, particularly machine learning and deep learning, has grown increasingly popular. These technologies primarily enhance the speed of analysis and the accuracy of emotion classification (He et al., 2022; Begum et al., 2023; Villegas-Ch et al., 2023; Yugal et al., 2023). Although artificial intelligence has seen significant advancements recently, various models have also been employed for speech emotion recognition to explore the relationship between emotions and academic performance (Dehbozorgi and Kunuku, 2023).

In this sense, numerous studies have performed detailed analyses to uncover the relationship between students' expressed emotions and their academic performance (Dehbozorgi and Kunuku, 2023; Yugal et al., 2023). Positive emotions like relief and satisfaction are strongly correlated with higher grades, suggesting that students experiencing these emotions typically achieve better academically. Conversely, negative emotions, such as frustration, are negatively correlated with academic performance, indicating that students experiencing these emotions often face academic challenges.

Positive emotional states have been associated with greater success in completing class activities on time and better overall performance, underscoring the importance of emotional well-being in academic settings (Dehbozorgi and Kunuku, 2023).

These technologies have been instrumental in identifying the impact of emotions on learning outcomes, linking positive emotions to improved cognitive processes and engagement (He et al., 2022; Villegas-Ch et al., 2023; Yugal et al., 2023).

3.2 Technology integration and learning outcomes

In the current educational landscape, integrating technology is key to enhancing learning outcomes. Blending technological tools with traditional teaching methods has created new opportunities to enrich the educational experience and foster skill development in students. The effectiveness of this integration is evident in its adaptability to various learning styles, its ability to boost student engagement, and its role in providing access to global educational resources.

In the context of our research on using artificial intelligence to assess emotions in learning environments, it is essential to understand how this synergy between technology and learning outcomes can improve the educational process and make achieving learning objectives more effective and meaningful.

Technologies such as AI, machine learning, and deep learning are employed to expedite emotion analysis and enhance classification accuracy in educational settings (He et al., 2022; Dehbozorgi and Kunuku, 2023). The integration of artificial intelligence (AI) into the management of emotions within education marks a significant advancement in modern teaching methods. Research has shown that machine learning techniques can reliably identify a range of human emotions, including happiness, anger, sadness, and calmness (Ramirez and Vamvakousis, 2012).

This ability can significantly enhance teaching by providing a deeper, more personalized understanding of students' emotional states. Such insights allow for the customization of teaching strategies to better address individual needs. Effectively applying AI in managing educational emotions can not only boost students' overall well-being but also foster a more inclusive and empathetic learning environment.

Improved emotional recognition from "EEG signals" can be enhanced by integrating deep learning with shallow machine learning techniques, which holds promising applications in human-computer interaction (Islam et al., 2021). This development signifies a major research advancement, recognizing deep learning's ability to extract complex features from EEG signals and the role of shallow machine learning in providing a clearer, more interpretable analysis. Combining these approaches creates a synergistic effect, enhancing the detection and understanding of emotions from EEG signals. Such advancements could lead to innovative applications in human-computer interaction, resulting in more intuitive and adaptive interfaces that align with users' emotions and needs.

Artificial intelligence-based educational models, like "MAPLE," are poised to positively influence student motivation and engagement in e-learning environments by catering to their emotional needs (Mehigan and Pitt, 2019). This underscores the value of adaptability and personalization in these systems, allowing for more targeted responses to learners' emotional states. By incorporating artificial intelligence, educational environments become more responsive and empathetic, thereby enhancing student engagement and satisfaction.

Another significant development is the emergence of affective computing and sentiment analysis. These fields utilize human-computer interaction, information retrieval, and multimodal signal processing to analyze sentiments from online social data, providing valuable insights for educational sciences (Cambria, 2016; Cambria et al., 2017). These advancements facilitate a deeper understanding of emotional experiences in digital settings, which can inform both online and offline educational strategies. Integrating these emotional analytics into education enhances the customization of teaching methods and curriculum design to better meet students' emotional needs, promoting more effective and meaningful learning experiences.

3.2.1 Emotionally intelligent e-learning

Emotion recognition is vital for understanding the influence of emotions on peer interactions, academic performance, and engagement in online and virtual learning environments (Standen et al., 2020; He et al., 2022; Dehbozorgi and Kunuku, 2023).

Emotions play a crucial role in human interaction and decision-making processes. EEG signals provide an accessible, inexpensive, portable, and user-friendly means to identify emotions (Alarcao and Fonseca, 2017). This technology is highly valued for its real-time analysis capabilities of emotional states. The portability and ease of use of EEG devices make them particularly suitable for educational applications, offering new possibilities for enhancing communication, well-being, and decision-making at both individual and societal levels.

Transfer learning approaches, which utilize networks pretrained on other tasks, have proven highly effective in facial emotion recognition within human-computer interaction, achieving an impressive average accuracy of 96% (Chowdary et al., 2023). This method leverages the existing knowledge embedded in neural network models to enhance the detection of emotional expressions in digital settings. The high accuracy of these approaches lays a strong foundation for developing advanced human-computer interaction systems, which can enhance online learning experiences by providing more accurate and nuanced emotional feedback.

Chao et al. (2019) introduced a deep learning framework that employs a multiband feature matrix and a CapsNet model to improve emotion recognition from multi-channel EEG signals, outperforming common models. This innovation underscores the importance of advancing deep learning techniques to increase the accuracy and efficiency of emotion recognition in educational settings. By integrating multiple EEG channels and utilizing the generalization capabilities of CapsNet models, this framework sets a new standard for detecting emotional states, significantly impacting our understanding of emotions in academic performance and engagement in online and virtual learning environments.

The novel deep learning model (ERDL), which combines graph convolutional neural networks and LSTMs, has achieved superior classification accuracy for emotion recognition from EEG signals compared to current state-of-the-art methods (Yin et al., 2021). This advancement underscores the effectiveness of integrating various deep learning techniques to enhance emotional recognition in brain signals. By combining the capability to model complex relationships in graph-like data with the ability to handle temporal sequences through LSTMs, the ERDL model emerges as a potent tool for deciphering emotions via EEG signals. This improvement in classification accuracy is crucial for designing more effective educational interventions tailored to the emotional needs of students.

Development began in 2014 of a technique using convolutional neural networks that effectively learns emotion-relevant features from speech, maintaining stable and robust performance even in complex environments (Mao et al., 2014). This study demonstrates the power of convolutional neural networks in extracting distinct emotion-related features from speech, enabling precise and reliable recognition of emotional expressions across various settings. The consistent and robust performance of this method supports its potential for practical applications, including enhancing human-computer interaction in virtual and online educational settings.

Furthermore, research indicates that students' understanding of emotions correlates positively with their academic performance, peer acceptance, and school adaptation, especially among children from middle-class families (Voltmer and von Salisch, 2017). This finding highlights the importance of emotional intelligence in the educational and social contexts of students, influencing various aspects of their development. The ability to understand and manage emotions not only affects academic success but also enhances the quality of interpersonal relationships and adaptability in school settings. Additionally, the variation in these associations across different socioeconomic backgrounds emphasizes the need for equitable attention to emotional development within education.

3.2.2 Emotionally intelligent e-learning systems and adaptive learning systems

Emotionally Intelligent E-learning Systems (EIES) and adaptive learning systems are transforming learning experiences by providing personalized educational environments (Ninaus et al., 2019; He et al., 2022; Dehbozorgi and Kunuku, 2023).

The Emotionally Intelligent E-Learning System (EIES), based on the Bayesian Network model, accurately predicts students' emotions during online learning sessions, enhancing the quality of virtual education (Daouas and Lejmi, 2018). This innovation underscores the importance of incorporating emotional intelligence into online learning environments. By predicting emotions, EIES can dynamically tailor the delivery of educational content, provide personalized feedback, and offer emotional support resources when necessary. This capability significantly enriches the online learning experience, creating a more responsive and engaging educational environment.

Additionally, it has been demonstrated that artificial intelligence techniques can enhance adaptive e-learning platforms by creating detailed learner profiles and models, which in turn improve the learning process and reduce uncertainty (Colchester et al., 2017). This advancement highlights the crucial role of artificial intelligence in personalizing online education by enabling systems to adapt dynamically to individual learner needs. Advanced algorithms help these platforms identify specific learning patterns, preferences, and challenges of each student, thereby facilitating the delivery of relevant and effective educational content. This adaptive capability significantly improves the learning experience, fostering a more responsive and student-centered educational environment.

A cloud-based adaptive learning system has proven effective in integrating mobile devices into the classroom environment, providing real-time feedback and context-aware content adaptation, leading to significant improvements in student performance and achievement (Nedungadi and Raman, 2012).

This approach demonstrates the potential of mobile technology and cloud computing to enhance the classroom learning experience by offering greater flexibility and personalization of educational content. By leveraging mobile devices like tablets or smartphones, adaptive systems can deliver instant feedback and tailor content to individual needs and learning contexts, thus boosting the overall effectiveness of the educational process and encouraging student participation and engagement.

Adaptive learning technologies, which tailor instruction to align with students' personal interests, have demonstrated the potential to enhance performance and learning outcomes (Walkington, 2013). This finding emphasizes the importance of customizing educational content to match individual preferences to optimize the learning process. Adaptive algorithms analyze students' learning patterns and interests, allowing systems to present relevant content and challenges that maintain their motivation and engagement. This personalized approach promotes more active participation and a deeper understanding of the material, ultimately leading to improved academic performance and more positive learning outcomes.

Personalized adaptive learning, facilitated by intelligent learning environments, combines personalized and adaptive learning strategies, making adaptive adjustments to teaching approaches based on individual characteristics, performance, and personal development (Peng et al., 2019). This integration of educational methods offers a comprehensive and effective solution tailored to the unique needs of each learner. By merging personalized educational content with dynamic adaptations in teaching methodology, it creates an educational environment that continually adjusts to the abilities, interests, and preferences of students. This not only maximizes each individual's learning potential but also enhances engagement and motivation toward the educational process.

In summary, the integration of Emotionally Intelligent E-Learning Systems (EIES) and adaptive learning systems significantly enhances the educational experience by providing personalized environments that dynamically adapt to the emotional and learning needs of students. This synergy between advanced technologies and contemporary educational methodologies supports the accurate prediction of students' emotions during online learning sessions and the real-time adaptation of content and teaching strategies. Collectively, these advancements underscore the transformative role of technology in education, promoting more effective, inclusive, and student-centered learning environments.

3.2.3 Positive emotional states and academic performance

Positive emotional states correlate strongly with improved academic performance and increased engagement in online learning environments. Students' expressed emotions, such as relief, satisfaction, and frustration, are directly linked to their academic outcomes, illustrating the significant impact of emotions on learning results (Ninaus et al., 2019; He et al., 2022; Dehbozorgi and Kunuku, 2023).

Academic emotions, ranging from anxiety to other emotional states, have a significant impact—both positive and negative—on students' motivation, learning strategies, self-regulation, and academic performance (Pekrun et al., 2017). This study illustrates how different emotional states can affect various aspects of academic performance and student engagement. Anxiety, for instance, can impede motivation and self-regulation, while positive emotions can enhance learning strategies and promote greater engagement with study materials. Understanding the interaction between emotions and academic performance underscores the importance of creating an educational environment that promotes positive emotional states and provides support to effectively manage negative emotions.

Moreover, students' emotions, whether negative or positive, significantly influence their academic performance, with cognitive processes and effortful control playing a moderating role in this relationship (Valiente et al., 2012). This study highlights the complex interplay between emotions and cognitive processes in the educational context, noting how effortful control can modulate the impact of emotions on academic performance. Positive emotions can enhance performance by promoting greater motivation and engagement, while negative emotions may hinder performance by interfering with attention and memory. The role of effortful control suggests that emotional and cognitive regulation strategies can mitigate the negative effects of adverse emotions and amplify the benefits of positive emotions on academic performance.

Positive emotions, such as enjoyment and pride, are positively associated with mathematics achievement, while negative emotions, such as anger, anxiety, shame, boredom, and hopelessness, have a negative correlation (Pekrun et al., 2017). This emphasizes the importance of emotions in the academic context and their differential impact on student performance. Positive emotions can boost motivation and readiness for learning, whereas negative emotions can generate distractions and cognitive blocks. These findings highlight the need to foster an educational environment that encourages positive emotions and provides effective strategies to manage negative emotions, aiming to improve both academic performance and student well-being.

Positive emotions also promote academic performance in college students when mediated by self-regulated learning and motivation (Mega et al., 2014). This study demonstrates that positive emotions not only directly influence academic performance but also interact with internal processes such as self-regulation of learning and motivation to enhance educational outcomes. Positive emotions can increase perseverance, attention, and the effectiveness of self-regulated learning strategies, improving comprehension and retention of academic material. Additionally, these emotions can reinforce intrinsic motivation and disposition toward learning, leading to deeper and more sustained engagement with the educational process.

Lastly, positive emotions positively influence problem-solving patterns by engaging students in self-regulatory activities, whereas negative emotions result in less variety of search activities and fewer regulatory activities (Zhou, 2013). This study shows how emotions can shape the way students approach academic challenges and handle complex problems. Positive emotions encourage active exploration, creativity, and cognitive flexibility, leading to a wide range of problem-solving strategies and greater solution-finding effectiveness. Conversely, negative emotions can restrict students' ability to think creatively and seek alternative solutions, resulting in less diversity in problem-solving strategies and approaches. These findings underscore the importance of promoting a positive emotional classroom environment to foster the development of effective problem-solving skills and self-regulation in students.

Positive emotions are associated with higher academic performance as they enhance psychological capital, which includes elements like efficacy, hope, optimism, and resilience (Carmona-Halty et al., 2019). This association supports the idea that positive emotional states correlate with more effective cognitive processes, better academic outcomes, and greater engagement in online learning environments.

The emotional climate of the classroom also has a significant impact on academic achievement, fostering greater student participation across all grade levels and genders (Reyes et al., 2012). This supports the view that a positive emotional environment in the classroom is crucial for academic success as it enhances student engagement and involvement in the educational process, thereby improving learning outcomes.

From a broader perspective, the TPACK framework emphasizes the effective integration of technological, pedagogical, and disciplinary content to enhance learning outcomes (Alemán-Saravia and Deroncele-Acosta, 2021). In this context, attention-based convolutional recurrent neural networks (ACRNN) are notable for their ability to accurately extract discriminative features from EEG signals, improving emotion recognition over other methods (Tao et al., 2020). The integration of AI into educational design influences learning outcomes directly, increasing motivation, self-efficacy, and the effectiveness of cognitive learning strategies within learning communities (Stefanou and Salisbury-Glennon, 2002). Moreover, AI's role in detecting students' emotions not only enhances productivity and academic performance (AlZu'bi et al., 2022) but also streamlines teaching practices by allowing educators to monitor emotional states and provide targeted feedback that positively affects learning outcomes (Deniz et al., 2019).

Additionally, AI can automate assessment-related decisions, optimizing the effectiveness of computerized formative assessments to enhance student learning (Shin et al., 2022) and predict student performance with high accuracy, enabling early interventions and ensuring equitable quality education (Jokhan et al., 2022). AI applications in education are diverse, including profiling, assessment, adaptive systems, personalization, and intelligent tutoring systems (Zawacki-Richter et al., 2019).

Furthermore, course-related discussions and interactions among students are shown to positively influence grades more than non-course-related topics, underscoring the importance of emotional engagement in learning.

Scientific evidence indicates that discussions within online learning management systems can enhance student engagement, improve the content and quality of work, and lead to better learning outcomes (King et al., 2021). Aligning with this, another study suggests that highly interactive online courses—marked by substantial student-to-student and student-to-instructor interactions—are perceived more favorably in terms of engagement and learning outcomes compared to less interactive group courses and discussions (Tsai et al., 2021). Additionally, peer discussions have been shown to enhance student performance on conceptual questions in class, fostering greater understanding and improved accuracy, even when none of the students initially know the correct answer (Smith et al., 2009).

Therefore, classroom interaction and discussion are crucial factors for learning, and promoting these should be a priority within educational systems. The integration of AI can support this goal, as AI techniques can effectively identify significant contributions and patterns in students' electronic discussions. This capability assists teachers in fostering productive discussions and enhancing learning (McLaren et al., 2010). Furthermore, AI can also be utilized to develop students' skills in complex interpersonal behaviors, such as effective listening, teamwork, and communication (Hoffmann-Longtin et al., 2018).

3.3 Special education and assistive technology

Information and communication technologies (ICT) and assistive technologies are vital for helping students, both with and without disabilities, to recognize their emotions and enhance their learning. These technologies are particularly crucial in removing barriers for children with learning difficulties. Research shows that ICT applications can create inclusive learning environments and provide essential support for students with learning challenges (Standen et al., 2020; Begum et al., 2023).

Educators can utilize ICT and assistive technologies to customize learning experiences based on the emotional needs of individual students, thereby improving their engagement and overall learning outcomes. The incorporation of these technologies not only aids in emotion recognition but also establishes a supportive learning environment that promotes both emotional well-being and academic success for students with diverse learning needs (Standen et al., 2020; Begum et al., 2023).

For students with special needs, mobile learning provides greater accessibility and richer learning experiences, presenting a valuable alternative to traditional assistive devices. This mode of learning enables students with diverse needs, including those with disabilities, to engage in more adaptive and personalized learning (Standen et al., 2020).

By leveraging mobile technologies, educators can create inclusive learning environments that cater to individual learning styles and preferences, thereby enhancing student engagement and academic outcomes. Mobile devices are portable and versatile, making learning more accessible and convenient for students with special needs, allowing them to interact with educational content in a manner that best suits their specific requirements.

The integration of mobile learning not only improves accessibility but also enables students with special needs to participate more actively in their learning process, fostering independence and encouraging self-directed learning (Standen et al., 2020).

Overall, the strategic use of ICT and assistive technologies not only facilitates the identification of emotions but also nurtures an environment conducive to learning, supporting the emotional well-being and academic achievements of students with diverse learning needs.

3.4 Affective computing

Affective computing is an interdisciplinary field that develops systems capable of recognizing, interpreting, and responding to human emotions.

This field of research has significant relevance in education because emotions are critical to the learning process and to creating meaningful educational experiences.

Understanding students' emotions allows educators to tailor their teaching methods to more effectively meet individual needs and foster a positive and stimulating learning environment. In the context of our research on integrating artificial intelligence to assess emotions in learning environments, we investigate how advancements in affective computing can enhance the assessment of students' emotional experiences and improve learning outcomes.

3.4.1 Affective computing as an area of AI in emotional management

As explained by Villegas-Ch et al. (2023), affective computing is a branch of Artificial Intelligence (AI) developed to enable computer systems to interact with humans effectively. This interaction is facilitated through computer vision techniques and machine learning algorithms. The primary goal is to produce a system that can elicit effective responses from users. Affective computing is interdisciplinary and consists of four main research areas: (1) Analysis and characterization of affective states, (2) Automatic recognition of affective states through facial expressions, linguistic features, posture, gaze tracking, and heart rate, among others, (3) System adaptation to respond appropriately to the users' affective states, and (4) Design of avatars that display suitable affective responses for better user interaction.

Emotion recognition via facial expressions, often referred to as facial expression recognition, is a widely addressed topic within the field of affective computing (González-Hernández et al., 2018). By recognizing facial expressions, educators can offer more personalized responses, provide emotional support when needed, and promote a more empathetic and student-centered learning environment. Thus, the ability to recognize emotions through facial expressions is essential for enhancing the quality and effectiveness of education.

3.4.2 Optimizing learning through affective computing

The assertion that affective computing is essential to intelligent learning systems is strongly supported by the growing recognition that emotions are integral to cognitive and decision-making processes. Research such as Shobana and Kumar (2021) demonstrates that emotions significantly influence perception and learning. Integrating affective computing into educational systems enables a more precise and personalized response to students' emotional needs, thereby enhancing the effectiveness of the teaching and learning process. Recognizing and responding to students' emotions opens new possibilities for creating empathetic and effective learning environments, ultimately fostering deeper and more meaningful learning experiences.

One of the key challenges of affective computing is the automatic detection and classification of users' emotional reactions to learning materials (Ninaus et al., 2019). This capability is crucial in education for several reasons. Firstly, it allows educational systems to adapt personally to the emotional needs of students, enhancing the learning experience and fostering a more responsive and empathetic environment. Additionally, it enables the early identification of potential emotional difficulties that may impact academic performance, allowing for timely educational interventions.

There are online learning platforms that utilize affective computing principles to accurately identify six fundamental emotions: happiness, disgust, anger, surprise, sadness, and fear (Aly et al., 2023). Recognizing and addressing this range of emotions allows educational strategies to be more contextualized and effective. For example, detecting happiness can lead to the reinforcement of student achievements, maintaining a motivating environment. Recognizing disgust can help avoid content that triggers negative reactions, thus enhancing the learning experience. By identifying anger, platforms can provide additional support to help students overcome challenges and stay motivated. Surprise can indicate moments of insight, which can be leveraged to deepen understanding. Recognizing sadness is essential for providing emotional support, while identifying fear can signal the need for psychoeducational interventions to manage stress and ensure effective learning. Overall, these capabilities facilitate a more adaptive and emotionally-aware approach, promoting a more inclusive and effective educational environment.

Research has extensively explored methods and models for affect detection systems capable of analyzing conventional modalities such as facial expression, voice, body language and posture, physiology, brain imaging, and multimodal systems. This research connects human emotions to learning, organizing them into four quadrants—curiosity, confusion, frustration, hope—with emotions on the horizontal axis and learning on the vertical axis (Kort et al., 2001).

Affective computing offers multiple benefits when integrated with artificial intelligence for emotion recognition. It has been shown to enhance e-learning applications by detecting and responding to the emotions of learners, potentially improving the learning process (Thompson and McGill, 2012). It can adjust the mood of learners to create a more effective learning environment (Chen and Lee, 2012), recognize emotions from speech using neural networks (Zhang et al., 2007), monitor students' behavior to gauge their attention and engagement levels, and support effective learning processes (Bevilacqua et al., 2009). Additionally, it can boost motivation and satisfaction in game-based adaptive learning systems (Tsai et al., 2012). In the context of game-based learning, one study shows that adaptive gamification—which combines artificial intelligence, gamification, and educational data mining—has a positive impact on student engagement and learning performance (Daghestani et al., 2020).

3.4.3 Advantages of affective computing as an intelligent educational system

The primary aim of affective computing is to develop an "intelligent" computer system capable of sensing, recognizing, understanding, and intelligently responding to human emotions in a timely and friendly manner (He et al., 2022). Affective computing is an interdisciplinary field dedicated to creating systems and technologies that can recognize, interpret, process, and respond to human emotions. This field strives to equip machines with the capability to comprehend and mimic human emotional intelligence, utilizing a variety of data sources like facial expressions, tone of voice, handwriting patterns, and other physiological indicators to ascertain a person's emotional state.

In educational settings, when integrated with artificial intelligence, affective computing can personalize teaching by adapting content to align with students' emotions and individual needs (Kratzwald et al., 2018; Marín-Morales et al., 2018, 2020; Arnau-González et al., 2021; Li S. et al., 2021; Li W.-C. et al., 2021; Wang et al., 2022). As such, affective computing is a crucial element in the application of artificial intelligence in emotional management within educational environments. It allows AI to interpret facial expressions, voice tones, and other emotional cues, providing insights into students' emotional states. This capability not only facilitates the personalized adaptation of educational content but also enables the early identification of potential emotional challenges that may impact academic performance. Overall, the integration of affective computing into educational emotional management not only enhances the effectiveness of learning environments but also supports the emotional well-being of students, fostering a more supportive and responsive educational setting.

Several advantages and contributions of affective computing are recognized in recent research. For example, the Probability and Integrated Learning (PIL) algorithm effectively recognizes high-level human emotions, offering potential benefits for affective computing (Jiang et al., 2020). Additionally, fuzzy cognitive maps can accurately predict artificial emotions, aiding in the design of affective decision-making systems within AI (Salmeron, 2012).

In the context of e-learning systems, affective computing involves using tools to recognize users' emotions and adapt educational systems accordingly (Jaques and Viccari, 2006). It has been demonstrated that affective computing can detect human attention levels using multimodal inputs such as webcams and mouse movements, potentially enhancing performance in intelligent e-learning applications (Li et al., 2016).

Broadly speaking, affective computing is a field within artificial intelligence that focuses on developing systems capable of recognizing, interpreting, processing, and simulating human emotions. It employs machine learning techniques, computer vision, natural language processing, and other disciplines to analyze and respond to the emotions expressed by users. The ultimate goal is to create systems that are more empathetic and can interact more naturally with people. Particularly in education, integrating affective computing is crucial for understanding and addressing student emotions, promoting a learning environment that is more personalized, effective, and attentive to emotional well-being.

4 Conclusion

The integration of assistive technology, information and communication technology (ICT), and artificial intelligence (AI) in educational settings has revolutionized the support available to students, particularly those with learning difficulties, in managing their learning and emotions. For children with diverse learning needs, AI-enhanced emotion detection, personalized learning experiences through ICT, and improved accessibility via assistive technology have significantly reduced learning barriers.

This research highlights the critical role of technology in enhancing emotion recognition, creating inclusive learning environments, and promoting academic success for all children. By employing these advanced tools, educators can develop customized learning plans, provide immediate feedback, and support both the academic and emotional development of students with and without disabilities.

This comprehensive approach to integrating AI, ICT, and assistive technology not only enhances emotional support but also equips students with the tools they need to actively participate in their education. Ultimately, this opens the door to a more successful and inclusive educational process.

4.1 Limitations

This study aimed to review and analyze the existing literature on the integration of artificial intelligence for evaluating emotions in educational environments. The review relied on articles sourced from specific academic databases, including PubMed, Web of Science, PsycINFO, and Scopus. While these databases are significant in the scientific community, it is crucial to note that this selection might have limited the inclusion of pertinent research published in other sources or in gray literature.

Additionally, a linguistic bias is acknowledged; the review covered articles in English, Spanish, and Portuguese, but research published in other languages was not considered. This restriction might have excluded studies that could provide valuable insights, affecting the geographical and cultural representativeness of the studies included in this analysis. Consequently, the generalizability of the findings to different educational and cultural contexts may be limited.

Lastly, we recognize an open access bias. Despite efforts to include a diverse range of academic sources and databases, some studies may be behind paywalls. This limitation could have excluded significant research, impacting the comprehensiveness of the review. Access was restricted to studies that were either open access or available through institutions with subscriptions. Therefore, caution is advised when interpreting the findings of this review, as they may not comprehensively reflect all the available research in the field of AI-driven emotional assessment in education.

4.1.1 Temporal limitation

The review findings may not fully represent the latest advances or developments in AI-driven emotional assessment within education, as the search was conducted up to the year 2023. Consequently, emerging technologies, methodologies, or ethical considerations may not be sufficiently covered, potentially limiting the relevance and applicability of the study's findings to current educational practices.

4.1.2 In terms of generalizability

While the review offers insights into the integration of AI for emotional assessment in educational settings, the findings may not be broadly applicable across diverse educational environments, student populations, and cultural contexts. Differences in educational infrastructure, resources, and practices among various regions or institutions could influence the feasibility and effectiveness of implementing AI-driven approaches to emotional assessment.

4.2 Ethical considerations

Although the study acknowledges the importance of ethical considerations in the development and implementation of AI technologies within education, the review itself does not delve into the ethical implications of using AI for emotional assessment. Further exploration of ethical frameworks, privacy concerns, and potential social impacts is needed to ensure responsible and equitable implementation of AI technologies within educational settings.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

AV: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. AD-A: Writing – original draft, Methodology, Formal Analysis, Data curation, Validation, Investigation, Writing – review & editing, Visualization, Conceptualization. JA: Writing – review & editing, Visualization, Validation, Supervision, Resources, Investigation, Conceptualization. AB: Supervision, Project administration, Writing – review & editing, Visualization, Methodology, Conceptualization. CL-G: Validation, Methodology, Writing – review & editing, Visualization, Supervision, Conceptualization. MM-G: Writing – review & editing, Visualization, Supervision, Resources, Project administration, Investigation, Conceptualization.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1387089/full#supplementary-material

References

Alanazi, S. A., Shabbir, M., Alshammari, N., Alruwaili, M., Hussain, I., and Ahmad, F. (2023). Prediction of emotional empathy in intelligent agents to facilitate precise social interaction. *Appl. Sci* 13:1163. doi: 10.3390/app13021163

Alarcao, S. M., and Fonseca, M. J. (2017). Emotions recognition using EEG signals: a survey. *IEEE Trans. Affect. Comput.* 10, 374–393. doi: 10.1109/TAFFC.2017.2714671

Alemán-Saravia, A. C., and Deroncele-Acosta, A. (2021). *Technology, pedagogy and content (tpack framework): Systematic literature review*, 104–111. IEEE.

Alm, C. O., Roth, D., and Sproat, R. (2005). *Emotions from text: Machine learning for text-based emotion prediction*, 579–586.

Aly, M., Ghallab, A., and Fathi, I. S. (2023). Enhancing facial expression recognition system in online learning context using efficient deep learning model. *IEEE Access* 11, 121419–121433. doi: 10.1109/ACCESS.2023.3325407

AlZu'bi, S., Abu Zitar, R., Hawashin, B., Abu Shanab, S., Zraiqat, A., Mughaid, A., et al. (2022). A novel deep learning technique for detecting emotional impact in online education. *Electronics* 11:2964. doi: 10.3390/electronics11182964

Arguel, A., Lockyer, L., Kennedy, G., Lodge, J. M., and Pachman, M. (2019). Seeking optimal confusion: a review on epistemic emotion management in interactive digital learning environments. *Interact. Learn. Environ.* 27, 200–210. doi: 10.1080/10494820.2018.1457544

Arnau-González, P., Katsigiannis, S., Arevalillo-Herráez, M., and Ramzan, N. (2021). "Artificial intelligence for affective computing: an emotion recognition case study" in *AI for emerging verticals: human-robot computing, sensing and networking*. eds. M. Z. Shakir and N. Ramzan, 29–44.

Assielou, K. A., Haba, C. T., Kadjo, T. L., Goore, B. T., and Yao, K. D. (2021). A new approach to modelling students' socio-emotional attributes to predict their performance in intelligent tutoring systems. *J. Educ. E Learn. Res.* 8, 340–348. doi: 10.20448/JOURNAL.509.2021.83.340.348

Begum, F., Neelima, A., and Valan, J. A. (2023). Emotion recognition system for E-learning environment based on facial expressions. *Soft. Comput.* 27, 17257–17265. doi: 10.1007/s00500-023-08058-3

Bevilacqua, L., Capuano, N., Cascone, A., Ceccarini, F., Corvino, F., D'Apice, C., et al. (2009). Advanced user interfaces for e-learning. *J. E Learn. Knowledge Soc.* 5, 91–99. doi: 10.20368/1971-8829/356

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intell. Syst.* 31, 102–107. doi: 10.1109/MIS.2016.31

Cambria, E., Das, D., Bandyopadhyay, S., and Feraco, A. (2017). Affective computing and sentiment analysis. *Pract Guide Sentiment Analysis*, 1–10. doi: 10.1007/978-3-319-55394-8_1

Carmona-Halty, M., Schaufeli, W. B., Llorens, S., and Salanova, M. (2019). Satisfaction of basic psychological needs leads to better academic performance via increased psychological capital: a three-wave longitudinal study among high school students. *Front. Psychol.* 10:2113. doi: 10.3389/fpsyg.2019.02113

Chaffar, S., Derbali, L., and Frasson, C. (2009). Inducing positive emotional state in intelligent tutoring systems. *Front. Artific. Intelligence Appl* 200, 716–718. doi: 10.3233/978-1-60750-028-5-716

Chaffar, S., and Frasson, C. (2010). "Using emotional coping strategies in intelligent tutoring systems" in *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 6095 LNCS(PART2)*. Berlin, Heidelberg: Springer, 285–287.

Chao, H., Dong, L., Liu, Y., and Lu, B. (2019). Emotion recognition from multiband EEG signals using CapsNet. *Sensors* 19:2212. doi: 10.3390/s19092212

Chen, G., and Lee, M. (2012). Detecting emotion model in e-learning system. In International conference on machine learning and cybernetics, 1686–1691

Chowdary, M. K., Nguyen, T. N., and Hemanth, D. J. (2023). Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Comput. & Applic.* 35, 23311–23328. doi: 10.1007/s00521-021-06012-8

Colchester, K., Hagras, H., Alghazzawi, D., and Aldabbagh, G. (2017). A survey of artificial intelligence techniques employed for adaptive educational systems within e-learning platforms. *J. Artif. Intelligence Soft Comput. Res.* 7, 47–64. doi: 10.1515/jaiscr-2017-0004

Daouas, T., and Lejmi, H. (2018). Emotions recognition in an intelligent elearning environment. *Interact. Learn. Environ.* 26, 991–1009. doi: 10.1080/10494820.2018.1427114

Daghestani, L., Ibrahim, L., Al-Towirgi, R., and Salman, H. (2020). Adapting gamified learning systems using educational data mining techniques. *Comput. Appl. Eng. Educ.* 28, 568–589. doi: 10.1002/cae.22227

de Oliveira, E. R., and Rodrigues, P. (2021). A review of literature on human behaviour and artificial intelligence: contributions towards knowledge management. *Electron. J. Knowl. Manag.* 19, 165–179. doi: 10.34190/ejkm.19.2.2459

Dehbozorgi, N., and Kunuku, M. (2023). Exploring the influence of emotional states in peer interactions on students' academic performance. *IEEE Trans. Educ.* 67, 405–412. doi: 10.1109/TE.2023.3335171

Deniz, S., Lee, D., Kurian, G., Altamirano, L., Yee, D., Ferra, M., et al. (2019). *Computer vision for attendance and emotion analysis in school settings*, 0134–0139. IEEE.

D'Mello, S., and Graesser, A. (2012). Language and discourse are powerful signals of student emotions during tutoring. *IEEE Trans. Learn. Technol.* 5, 304–317. doi: 10.1109/TLT.2012.10

Florea, A., and Kalisz, E. (2005). Embedding emotions in an artificial tutor. In Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC'05)

Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K. J., and Tørresen, J. (2018). Mental health monitoring with multimodal sensing and machine learning: a survey. *Pervasive Mobile Comput* 51, 1–26. doi: 10.1016/j.pmcj.2018.09.003

Gómez-León, M. I. (2022). Development of empathy through socioemotional artificial intelligence. *Papeles Psicol* 43, 218–224. doi: 10.23923/pap.psicol.2996

González-Hernández, F., Zatarain-Cabada, R., Barrón-Estrada, M. L., and Rodríguez-Rangel, H. (2018). Recognition of learning-centered emotions using a convolutional neural network. *J. Intelligent Fuzzy Syst.* 34, 3325–3336. doi: 10.3233/JIFS-169514

Gorga, D., and Schneider, D. K. (2009). "Computer-based learning environments with emotional agents" in *Handbook of research on synthetic emotions and sociable robotics: new applications in affective computing and artificial intelligence*, 413–441.

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., et al. (2019). Artificial intelligence for mental health and mental illnesses: an overview. *Curr. Psychiatry Rep.* 21, 1–18. doi: 10.1007/s11920-019-1094-0

He, T., Li, C., Wang, J., Wang, M., Wang, Z., and Jiao, C. (2022). An emotion analysis in learning environment based on theme-specified drawing by convolutional neural network. *Front. Public Health* 10:958870. doi: 10.3389/fpubh.2022.958870

Hoffmann-Longtin, K., Rossing, J. P., and Weinstein, E. (2018). Twelve tips for using applied improvisation in medical education. *Med. Teach.* 40, 351–356. doi: 10.1080/0142159X.2017.1387239

Islam, M. R., Moni, M. A., Islam, M. M., Rashed-Al-Mahfuz, M., Islam, M. S., Hasan, M. K., et al. (2021). Emotion recognition from EEG signal focusing on deep learning and shallow learning techniques. *IEEE Access* 9, 94601–94624. doi: 10.1109/ACCESS.2021.3091487

Jaques, P., and Viccari, R. (2006). Considering Students' emotions in computer-mediated learning environments. *Web-Based Intelligent E Learn Syst*, 122–138. doi: 10.4018/978-1-59140-729-4.CH006

Jiang, D., Wu, K., Chen, D., Tu, G., Zhou, T., Garg, A., et al. (2020). A probability and integrated learning based classification algorithm for high-level human emotion recognition problems. *Measurement* 150:107049. doi: 10.1016/j.measurement.2019.107049

Jokhan, A., Chand, A. A., Singh, V., and Mamun, K. A. (2022). Increased digital resource consumption in higher educational institutions and the artificial intelligence role in informing decisions related to student performance. *Sustain. For.* 14:2377. doi: 10.3390/su14042377

King, A. S., Taylor, J. B., and Webb, B. M. (2021). Promoting productive political dialogue in online discussion forums. *J. Polit. Sci. Educ.* 17, 724–750.

Kort, B., Reilly, R., and Picard, R. W. (2001). *An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion*. Madison, WI, USA: Proceedings IEEE International Conference on Advanced Learning Technologies, 43–46.

Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., and Prendinger, H. (2018). Deep learning for affective computing: text-based emotion recognition in decision support. *Decis. Support. Syst.* 115, 24–35. doi: 10.1016/j.dss.2018.09.002

Lamm, C., and Singer, T. (2010). The role of anterior insular cortex in social emotions. *Brain Struct. Funct.* 214, 579–591. doi: 10.1007/s00429-010-0251-3

Li, J., Ngai, G., Leong, H., and Chan, S. (2016). Multimodal human attention detection for reading from facial expression, eye gaze, and mouse dynamics. *ACM Sigapp Appl. Comput. Rev.* 16, 37–49. doi: 10.1145/3015297.3015301

Li, W.-C., Yang, C.-J., Liu, B.-T., and Fang, W.-C. (2021). A real-time affective computing platform integrated with AI system-on-Chip Design and multimodal signal processing system, 522–526. doi: 10.1109/EMBC46164.2021.9630979

Li, S., Zheng, J., Lajoie, S. P., and Wiseman, J. (2021). Examining the relationship between emotion variability, self-regulated learning, and task performance in an intelligent tutoring system. *Educ. Technol. Res. Dev.* 69, 673–692. doi: 10.1007/s11423-021-09980-9

Lin, Y.-J., Ding, S. Y., Lu, C.-K., Tang, T. B., and Shen, J.-Y. (2023). *Emotion prediction in music based on artificial intelligence techniques*, 405–406.

Malekzadeh, M., Mustafa, M. B., and Lahsasna, A. (2015). A review of emotion regulation in intelligent tutoring systems. *Educ. Technol. Soc.* 18, 435–445.

Malekzadeh, M., Salim, S. S., and Mustafa, M. B. (2014). Towards integrating emotion management strategies in intelligent tutoring system used by children. In Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, 100, 41–50

Mao, Q., Dong, M., Huang, Z., and Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* 16, 2203–2213. doi: 10.1109/TMM.2014.2360798

McLaren, B. M., Scheuer, O., and Mikšátko, J. (2010). Supporting collaborative learning and e-discussions using artificial intelligence techniques. *Int. J. Artif. Intell. Educ.* 20, 1–46.

Marín-Morales, J., Higuera-Trujillo, J. L., Greco, A., Guixeres, J., Llinares, C., Scilingo, E. P., et al. (2018). Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Sci. Rep.* 8. doi: 10.1038/s41598-018-32063-4

Marín-Morales, J., Llinares, C., Guixeres, J., and Alcañiz, M. (2020). Emotion recognition in immersive virtual reality: from statistics to affective computing. *Sensors* 20, 1–26. doi: 10.3390/s20185163

Mega, C., Ronconi, L., and De Beni, R. (2014). What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *J. Educ. Psychol.* 106:121. doi: 10.1037/a0033546

Mehigan, T., and Pitt, I. (2019). *Engaging learners through emotion in artificially intelligent environments*, 5661–5668.

Mohanan, R., Stringfellow, C., and Gupta, D. (2018). *An emotionally intelligent tutoring system*, 1099–1107.

Mohr, D. C., Zhang, M., and Schueller, S. M. (2017). Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu. Rev. Clin. Psychol.* 13, 23–47. doi: 10.1146/annurev-clinpsy-032816-044949

Nedungadi, P., and Raman, R. (2012). A new approach to personalization: integrating e-learning and m-learning. *Educ. Technol. Res. Dev.* 60, 659–678. doi: 10.1007/s11423-012-9250-9

Ninaus, M., Greipl, S., Kiili, K., Lindstedt, A., Huber, S., Klein, E., et al. (2019). Increased emotional engagement in game-based learning – a machine learning approach on facial emotion detection data. *Comput. Educ.* 142. doi: 10.1016/j.compedu.2019.103641

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int. J. Surg.* 88:105906. doi: 10.1016/j.ijsu.2021.105906

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., and Goetz, T. (2017). Achievement emotions and academic performance: longitudinal models of reciprocal effects. *Child Dev.* 88, 1653–1670. doi: 10.1111/cdev.12704

Peng, H., Ma, S., and Spector, J. M. (2019). Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment. *Smart Learn. Environ.* 6, 1–14.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., et al. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. *A product from the ESRC methods programme Version*, 1:b92.

Ramirez, R., and Vamvakousis, Z. (2012). Detecting emotion from EEG signals using the emotive epoc device. *Brain Informatics. BI 2012. Lecture Notes in Computer Science*. eds. F. M. Zanzotto, S. Tsumoto, N. Taatgen and Y. Yao (Berlin, Heidelberg: Springer)175–184.

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., and Salovey, P. (2012). Classroom emotional climate, student engagement, and academic achievement. *J. Educ. Psychol.* 104:700. doi: 10.1037/a0027268

Rodríguez, A. O. R., Riaño, M. A., García, P. A. G., Marín, C. E. M., Crespo, R. G., and Wu, X. (2020). Emotional characterization of children through a learning environment using learning analytics and AR-sandbox. *J. Ambient. Intell. Humaniz. Comput.* 11, 5353–5367. doi: 10.1007/s12652-020-01887-2

Rodriguez, P., Ortigosa, A., and Carro, R. M. (2012). Extracting emotions from texts in e-learning environments. 887–892.

Salmeron, J. (2012). Fuzzy cognitive maps for artificial emotions forecasting. *Appl. Soft Comput.* 12, 3704–3710. doi: 10.1016/j.asoc.2012.01.015

Samsonovich, A. (2020). Socially emotional brain-inspired cognitive architecture framework for artificial intelligence. *Cogn. Syst. Res.* 60, 57–76. doi: 10.1016/j.cogsys.2019.12.002

Sarrafzadeh, A., Hosseini, H. G., Fan, C., and Overmyer, S. P. (2003). *Facial expression analysis for estimating learner's emotional state in intelligent tutoring systems*. 336–337.

Shatte, A. B. R., Hutchinson, D. M., and Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychol. Med.* 49, 1426–1448. doi: 10.1017/S0033291719000151

Shin, J., Chen, F., Lu, C., and Bulut, O. (2022). Analyzing students' performance in computerized formative assessments to optimize teachers' test administration decisions using deep learning frameworks. *J. Comput. Educ.* 9, 71–91. doi: 10.1007/s40692-021-00196-7

Singh, K., Goel, N., Gupta, B., and Bansal, D. (2023). Emotion prediction through facial recognition using machine learning: a survey. In Proceeding of the 2023 International Conference on Computer Communication and Informatics, ICCCI 2023

Shobana, B. T., and Kumar, G. A. S. (2021). I-quiz: an intelligent assessment tool for non-verbal behaviour detection. *Comput. Syst. Sci. Eng.* 40, 1007–1021. doi: 10.32604/CSSE.2022.019523

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., et al. (2009). Why peer discussion improves student performance on in-class concept questions. *Science* 323, 122–124. doi: 10.1126/science.1165919

Standen, P., Brown, D., Taheri, M., Trigo, M., Boulton, H., Burton, A., et al. (2020). An evaluation of an adaptive learning system based on multimodal affect recognition for learners with intellectual disabilities. *Br. J. Educ. Technol.* 51, 1748–1765. doi: 10.1111/bjet.13010

Stefanou, C. R., and Salisbury-Glennon, J. D. (2002). Developing motivation and cognitive learning strategies through an undergraduate learning community. *Learn. Environ. Res.* 5, 77–97. doi: 10.1023/A:1015610606945

Sumithra, M., Buvaneswar, B., Jessica Judith, S., and Punitha, R. (2022). Innovation for better education system using artificial intelligence. *J. Cognit. Human Computer Interact.* 2, 19–28. doi: 10.54216/JCHCI.020103

Sun, Y., Li, Z., and Xie, J. (2013). A formal model of emotional pedagogical agents in intelligent tutoring systems. *Scopus*. 319–323.

Taub, M., Azevedo, R., Rajendran, R., Cloude, E. B., Biswas, G., and Price, M. J. (2021). How are students' emotions related to the accuracy of cognitive and metacognitive processes during learning with an intelligent tutoring system? *Learn. Instr.* 72. doi: 10.1016/j.learninstruc.2019.04.001

Tao, W., Li, C., Song, R., Cheng, J., Liu, Y., Wan, F., et al. (2020). EEG-based emotion recognition via channel-wise attention and self attention. *IEEE Trans. Affect. Comput.*

Thompson, N., and McGill, T. (2012). Affective tutoring systems: enhancing e-learning with the emotional awareness of a human tutor. *Int. J. Inf. Commun. Technol. Educ.* 8, 75–89. doi: 10.4018/jicte.2012100107

Tsai, T., Lo, H., and Chen, K. (2012). An affective computing approach to develop the game-based adaptive learning material for the elementary students. *ACM Int. Conference Proceed. Series*, 8–13. doi: 10.1145/2160749.2160752

Tsai, C.-L., Ku, H.-Y., and Campbell, A. (2021). Impacts of course activities on student perceptions of engagement and learning online. *Distance Educ.* 42, 106–125. doi: 10.1080/01587919.2020.1869525

Valiente, C., Swanson, J., and Eisenberg, N. (2012). Linking students' emotions and academic achievement: when and why emotions matter. *Child Dev. Perspect.* 6, 129–135. doi: 10.1111/j.1750-8606.2011.00192.x

Villegas-Ch, W. E., Garcia-Ortiz, J., and Sanchez-Viteri, S. (2023). Identification of emotions from facial gestures in a teaching environment with the use of machine learning techniques. *IEEE Access* 11, 38010–38022. doi: 10.1109/ACCESS.2023.3267007

Voltmer, K., and von Salisch, M. (2017). Three meta-analyses of children's emotion knowledge and their school success. *Learn. Individ. Differ.* 59, 107–118. doi: 10.1016/j.lindif.2017.08.006

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: the impact of relevant contexts on performance and learning outcomes. *J. Educ. Psychol.* 105:932. doi: 10.1037/a0031882

Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., et al. (2022). A systematic review on affective computing: emotion models, databases, and recent advances. *Inform. Fusion* 83-84, 19–52. doi: 10.1016/j.inffus.2022.03.009

Yang, F., and Zhen, X. (2014). Research on the Agent's behavior decision-making based on artificial emotion. *The Journal of Information and Computational Science* 11, 2723–2733. doi: 10.12733/jics20103533

Yin, Y., Zheng, X., Hu, B., Zhang, Y., and Cui, X. (2021). EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM. *Appl. Soft Comput.* 100:106954. doi: 10.1016/j.asoc.2020.106954

Yugal, L., Kaswan, S., Bhatia, B. S., and Sharma, A. (2023). IoT-based emulated performance evaluation NLP model for advanced learners in academia 4.0 and industries 4.0. *J. Intelligent Syst. Internet Things* 10, 63–75. doi: 10.54216/JISIoT.100206

Zhang, Q., Wang, Y., Wang, L., and Wang, G. (2007). Research on speech emotion recognition in E-learning by using neural networks method. *IEEE International Conference on Control and Automation* 2007, 2605–2608. doi: 10.1109/ICCA.2007.4376833

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *Int. J. Educ. Technol. High. Educ.* 16, 1–27.

Zhou, M. (2013). "I am really good at it" or "I am just feeling lucky": the effects of emotions on information problem-solving. *Educ. Technol. Res. Dev.* 61, 505–520. doi: 10.1007/s11423-013-9300-y

TYPE Original Research
PUBLISHED 29 July 2024
DOI 10.3389/fpsyg.2024.1357975

OPEN ACCESS

EDITED BY
Cristian López Raventós,
National Autonomous University of Mexico, Mexico

REVIEWED BY
Minhwan Jang,
Rabdan Academy, United Arab Emirates
Giada Lettieri,
Université Catholique de Louvain, Belgium

*CORRESPONDENCE
Jang-Han Lee
✉ clipsy@cau.ac.kr

RECEIVED 19 December 2023
ACCEPTED 04 July 2024
PUBLISHED 29 July 2024

CITATION
Lee M-S, Lee G-E, Lee SH and Lee J-H (2024) Emotional responses of Korean and Chinese women to Hangul phonemes to the gender of an artificial intelligence voice.
Front. Psychol. 15:1357975.
doi: 10.3389/fpsyg.2024.1357975

Emotional responses of Korean and Chinese women to Hangul phonemes to the gender of an artificial intelligence voice

Min-Sun Lee[1], Gi-Eun Lee[2], San Ho Lee[3] and Jang-Han Lee[1]*

[1]Department of Psychology, Chung-Ang University, Seoul, Republic of Korea, [2]Institute of Cultural Diversity Content, Chung-Ang University, Seoul, Republic of Korea, [3]Department of European Language and Cultures, Chung-Ang University, Seoul, Republic of Korea

Introduction: This study aimed to explore the arousal and valence that people experience in response to Hangul phonemes based on the gender of an AI speaker through comparison with Korean and Chinese cultures.

Methods: To achieve this, 42 Hangul phonemes were used, in a combination of three Korean vowels and 14 Korean consonants, to explore cultural differences in arousal, valence, and the six foundational emotions based on the gender of an AI speaker. A total 136 Korean and Chinese women were recruited and randomly assigned to one of two conditions based on voice gender (man or woman).

Results and discussion: This study revealed significant differences in arousal levels between Korean and Chinese women when exposed to male voices. Specifically, Chinese women exhibited clear differences in emotional perceptions of male and female voices in response to voiced consonants. These results confirm that arousal and valence may differ with articulation types and vowels due to cultural differences and that voice gender can affect perceived emotions. This principle can be used as evidence for sound symbolism and has practical implications for voice gender and branding in AI applications.

KEYWORDS

phoneme, arousal, valence, emotions, sound symbolism, artificial intelligence voice

Introduction

Artificial intelligence (AI), which can be considered the core technology of the Fourth Industrial Revolution, has various applications. In particular, speakers, an important function of AI, facilitate communication between humans and AI. However, a challenge remains whether AI speakers can evolve from providing basic convenience functions, such as weather notifications and alarm settings, to providing better human–machine interaction. Previous studies on emotion recognition and emotional expression of AI have been actively conducted for better comprehensive communication. Most studies aimed to determine whether AI could recognize and express human-like emotions based on a user's emotional state. However, communication is two-way in nature. Thus, it is important to consider not only the emotions conveyed by AI but also how humans perceive the messages delivered by AI speakers. In the context of marketing channels, when users of AI speakers experience positive rather than negative emotions with AI speakers, their preference for these devices increases (Jang and Ju, 2019).

Humans communicate and express and spread emotions through language. Notably, the rapid identification of emotional elements in sound stimuli of communication plays an important role in survival and adaptation (Ramachandran and Hubbard, 2001). According to sound symbolism, specific phonemes, which are the fundamental elements of sound in a language, convey meaning independently (Lowrey et al., 2003). The *Bouba–Kiki effect*, a good example of sound symbolism, refers to the phenomenon of how people associate round shapes when they hear "Bouba" and pointed shapes when they hear "Kiki" (Ramachandran and Hubbard, 2001; Maurer et al., 2006; Pejovic and Molnar, 2017). Similarly, the *gleam–glum effect* demonstrates that words containing /i;/, such as "gleam," are perceived as more positive emotions than words containing /Λ/, such as "glum" (Yu, 2021). However, little agreement exists on whether these effects are universally applicable, regardless of the native language or age.

The question of whether phonemes have sound symbolism remains unanswered (Slunecko and Hengl, 2006). Although some studies have indicated a common theme of sound symbolism, the results vary, which is likely because the phonemes can be classified into consonants and vowels. Recent studies have revealed that the Bouba–Kiki effect varies between Eastern and Western cultures (Chen et al., 2016) and can change depending on differences in native language (Styles and Gawne, 2017). These previous findings suggest that sound–shape mapping related to consonants may be influenced by individual perceptual style and linguistic experience (Rogers and Ross, 1975; Chen et al., 2016; Shang and Styles, 2017; Chang et al., 2021).

This study adopted the classification of consonants as plain, aspirated, and voiced consonants, which is a common method and is recognized to evoke similar emotional impressions in various languages, including English and Korean. However, sound–size mapping associated with vowels is a common phenomenon across cultures and languages because of its lack of sensitivity to cultural backgrounds or native languages (Shinohara and Kawahara, 2010; Hoshi et al., 2019; Chang et al., 2021).

Therefore, vowels were selected based on the symbolism of vowel sounds. For instance, in the early research on sound symbolism by Sapir (1929), experiments were conducted on the size symbolism of vowels /a/ and /i/ using meaningless words "mal" and "mil." Participants were asked to identify those words that referred to a large table and a small table. Approximately 80% of participants indicated that "mal" denoted a large table and "mil" denoted a small table. This suggests that /a/, when added to an existing word, conveys a soft feeling because it is a central and low vowel, indicating augmentation for distant or large objects or long durations. Conversely, /i/ is considered to represent close and small objects or short durations. These research findings highlight the influence of mouth shape during pronunciation. In terms of the dimension of the aperture, high vowels, such as /i/ or /u/, involve a smaller aperture, whereas low vowels, such as /a/, involve a larger aperture, potentially conveying different symbolic meanings (Shinohara and Kawahara, 2010). Based on this evidence, this study adopted three representative vowel types that can induce different states and constructed 42 combinations of consonants and vowels.

To date, research on sound symbolism has mainly focused on vowels rather than consonants because consonants cannot be pronounced without vowels; thus, the sound symbols of vowels have been considered greater than those of consonants (Aveyard, 2012). However, actual language cannot ignore the influence of consonants, and comparing only the differences in vowels can limit recognition of the emotional meaning. Therefore, considering the practicality of language, this study attempts to measure the emotional values of both vowels and consonants through classification according to the articulation method (Kim, 2019).

When evaluating the emotional values of stimuli, arousal and valence are the two most basic dimensions (Russell, 1980). Arousal is evaluated based on how exciting or arousing a stimulus is, that is, how calm it is, and valence is evaluated based on how pleasant or unpleasant a stimulus is. According to empirical studies, the arousal and valence dimensions are not independent of each other and exhibit a *U*-shaped relationship. Thus, unpleasant stimuli are considered more arousing than pleasant stimuli, and both unpleasant and pleasant stimuli are more arousing than neutral stimuli (Libkuman et al., 2007; Grühn and Scheibe, 2008). In general, negative emotional stimuli are considered to have a higher arousal value than positive or neutral stimuli (Ekman et al., 1983).

However, preferences for words that express emotions show differences, indicating that cultural differences can also occur between language and emotional meanings (Park et al., 2018). Thus, even in the same situation that evokes emotions, the terms cognitively interpreted and referred to differ across cultures, and depending on how emotional words are translated, they can have distinct meanings (Hahn and Kang, 2000). Furthermore, many studies have measured the properties of sounds, such as rough, soft, strong, or weak. However, because of the ambiguous nature of these adjectives and their lack of integration into each study, accurately classifying how people feel about sound is not possible. Considering these points, this study adopted universal emotions to distinguish between subjective emotional states. The six basic emotions, namely anger, disgust, fear, sadness, surprise, and happiness (Ekman and Oster, 1979), were used to measure subjective emotional states instead of relying on somewhat ambiguous emotional expressions (e.g., softness, strength, weakness, and sharpness) based on the degree of arousal and valence (Russell, 1980, 2003; Barrett, 2006a,b).

The effect of AI speakers on human emotion recognition may include variables such as the voice gender of AI speakers, as well as sound-shape and sound-size. Studies have demonstrated differences in preference for "themes" by voice gender (Kim and Yun, 2021) and in AI usage behavior based on human gender and experience (Ji et al., 2019; Obinali, 2019; Ernst and Herm-Stapelberg, 2020; Kim and Yun, 2021; Wang et al., 2021). For example, in "warm news" delivery, female voices are highly appreciated in terms of understanding, reliability, and favorability, whereas, for news with serious content, male voices are preferred (Kim and Yun, 2021). Thus, gender preferences for an AI speaker may differ according to the gender of the human listener. Although arousal and valence can be perceived in phoneme units, studies on the voice gender of AI speakers have not yet observed an effect of voice gender on phoneme units.

Recently, active research has been conducted on how emotions are expressed and recognized in literary works using AI-based natural language processing and machine learning techniques. In addition, studies and reflections on AI-generated speech and listeners' emotional responses have rapidly evolved in recent years (Val-Calvo et al., 2020). Particularly, with advancements in speech recognition technology, there is growing interest in exploring how the tone and expression used by AI when speaking can evoke emotional responses in listeners (Poon-Feng et al., 2014; Zheng et al., 2015). Such studies provide crucial insights into understanding the impact of AI speech technology on people's emotional responses, aiming to offer important insights for the effective development and application of this technology.

This study used a cultural comparison to explore the arousal and valence that people experience in response to Hangul phonemes according to the gender of an AI speaker. For this purpose, the most basic unit, the Hangul phoneme, was used as an experimental stimulus to evaluate arousal and valence in Korean and Chinese women who can speak Korean. This study aimed to examine the cultural differences in arousal and valence using the articulation method, vowels, and the gender of the AI speaker.

TABLE 1 Psychometric characteristics of the participants.

Measure	Korean (n = 68)	Chinese (n = 68)	t (P-value)
PANAS-P	27.32 ± 5.26	28.56 ± 6.23	−1.25 (0.21)
PANAS-N	25.07 ± 5.69	26.97 ± 7.75	−1.63 (0.11)
STAI-T	47.29 ± 10.20	46.82 ± 6.77	0.32 (0.75)
STAI-S	46.68 ± 10.41	46.57 ± 6.93	0.068 (0.95)
CES-D	20.01 ± 10.83	18.15 ± 9.62	1.06 (0.29)

N = 136; all correlations are considered not statistically significant at $P < 0.01$. Mean ± standard deviation; PANAS-P, Positive and Negative Affect Schedule Scale-Positive; PANAS-N, Positive and Negative Affect Schedule Scale-Negative; STAI-T, State-Trait Anxiety Inventory-Trait Anxiety; STAI-S, State-Trait Anxiety Inventory-State Anxiety; CES-D, Center for Epidemiologic Studies Depression Scale.

Materials and methods

Participants and procedure

Using G*Power 3.1.9.7 (the University of Düsseldorf, Düsseldorf, Germany), a power analysis was conducted with an effect size of 0.25, an alpha error probability of 0.05, a power of 0.80, and the number of groups set to 4. The analysis showed that the minimum sample size required was 128 participants (42 participants per condition). In total, 136 participants were recruited (136 women; $M_{\text{age}} = 27.19$ years, $SD = 4.30$) from a university bulletin board in South Korea. The sample consisted of 68 Korean and Chinese participants who were randomly assigned to one of two conditions in a between-subjects design with voice gender (man or woman). All participants were informed that they had been recruited for a psychological experiment measuring emotions for phonemes and that all experimental processes would be conducted online. The study was limited to women to present the differences in variables, considering that women generally exhibit greater emotional responsiveness than men. The inclusion criteria for selecting participants were as follows: women who were (1) over the age of 20 years, (2) of Korean or Chinese nationality, and (3) able to speak Korean. Before taking part in the experiment, all participants provided informed consent and were informed that they could stop the experiment at any time. Each participant received $20 for their participation.

Measures

To control for emotional variables, the participants were asked to complete a questionnaire, described below. No significant differences were found in psychological characteristics between groups (Table 1).

TABLE 2 Forty-two Hangul phonetic values.

Articulation	Code	Corner vowels		
		ㅏ /a/	ㅜ /u/	ㅣ /i/
Lenis	ㄱ /g, k/	가 /ga, ka/	구 /gu, ku/	기 /gi, ki/
	ㄷ /d, t/	다 /da, ta/	두 /du, tu/	디 /di, ti/
	ㅂ /b, p/	바 /ba, pa/	부 /bu, pu/	비 /bi, pi/
	ㅅ /s/	사 /sa/	수 /su/	시 /si/
	ㅈ /j/	자 /ja/	주 /ju/	지 /ji/
	ㅎ /h/	하 /ha/	후 /hu/	히 /hi/
Aspirated	ㅊ /ch/	차 /cha/	추 /chu/	치 /chi/
	ㅋ /k/	카 /ka/	쿠 /ku/	키 /ki/
	ㅌ /t/	타 /ta/	투 /tu/	티 /ti/
	ㅍ /p/	파 /pa/	푸 /pu/	피 /pi/
Voiced	ㄴ /n/	나 /na/	누 /nu/	니 /ni/
	ㄹ /l, r/	라 /la, ra/	루 /lu, ru/	리 /li, ri/
	ㅁ /m/	마 /ma/	무 /mu/	미 /mi/
	ㅇ /ng/	아 /a/	우 /u/	이 /i/

English Notation for Articulation in Korean.

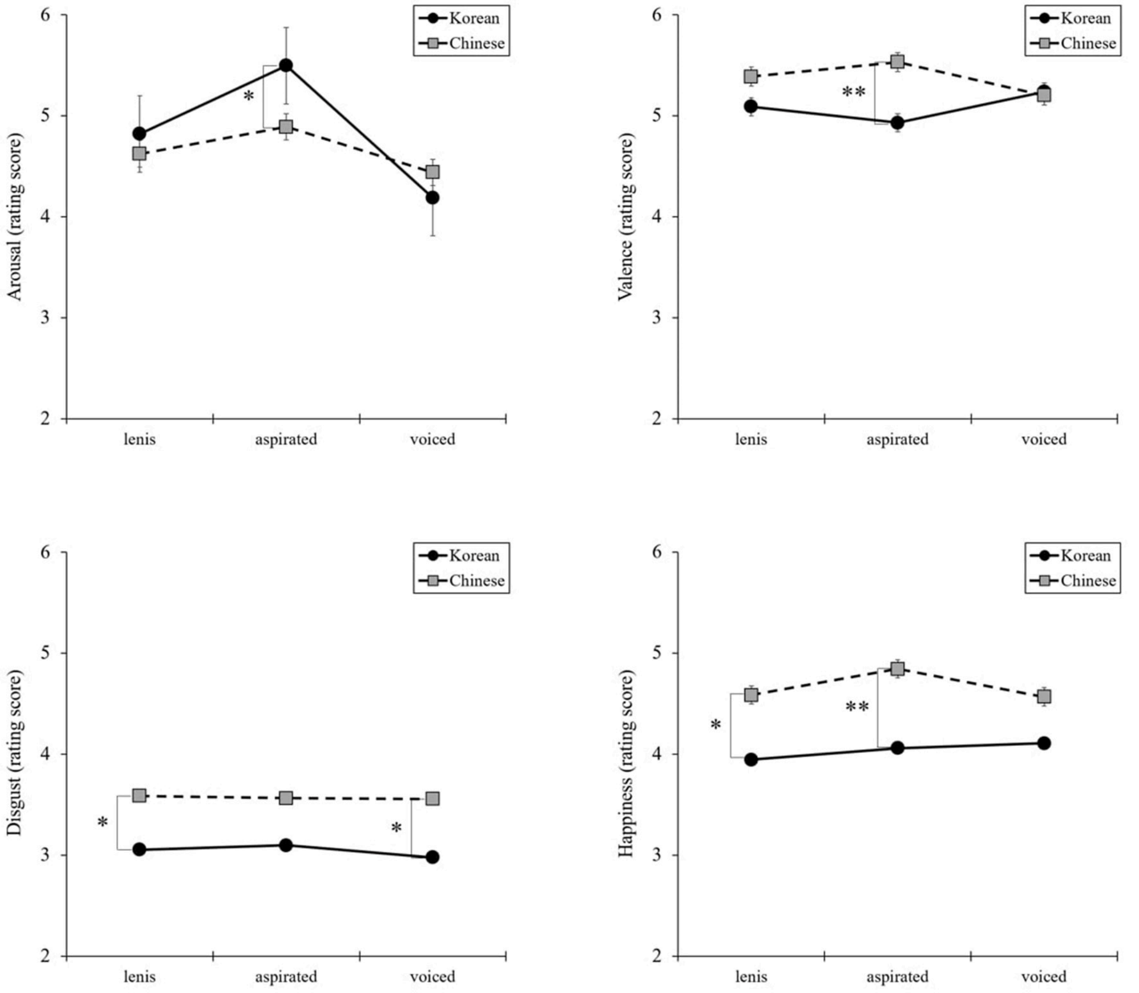


FIGURE 1
Interaction effects between the three types of articulation and nationality in arousal, valence, and basic emotions: disgust and happiness. *$P < 0.05$, **$P < 0.01$.

Positive and Negative Affect Schedule Scale

The Korean version (K-PANAS; Lee et al., 2003) and the Chinese version (C-PANAS; Huang et al., 2003) of the PANAS were used to evaluate the positive and negative affects. The PANAS comprises 20 items, with 10 evaluating expectations for positive affect (PANAS-P) and 10 evaluating expectations for negative affect (PANAS-N). Participants were asked to rate their responses on a 5-point Likert scale, where 1 indicates "very slightly or not at all" and 5 indicates "extremely." The higher the score, the higher the levels of positive and negative affect. Cronbach's alpha values were 0.60 and 0.83 for the K-PANAS-P and C-PANAS-P, respectively, and 0.61 and 0.85 for the K-PANAS-N and C-PANAS-N, respectively.

State-Trait Anxiety Inventory

The Korean version (K-STAI; Kim and Shin, 1978) and the Chinese version (C-STAl; Tsoi et al., 1986) of the STAI were used to measure state and trait anxiety (Spielberger et al., 1970). This scale consists of 40 items, with 20 measuring "trait anxiety (STAI-T)," and 20 measuring "state anxiety (STAI-S)." Participants were asked to rate on a 4-point Likert scale, where 1 indicates "not at all" and 4 indicates "very much so." The higher the score, the more intense or more often an individual felt anxious. Cronbach's alpha values of the K-STAI-T and K-STAI-S were 0.91 and 0.92, respectively, and those of the C-STAI-T and C-STAI-S were 0.75 and 0.74, respectively.

Center for Epidemiologic Studies Depression Scale

The Korean version (K-CES-D; Chon et al., 2001) and the Chinese version (C-CES-D; Chi and Boey, 1993) of the CES-D were used to measure the baseline for depressive mood in participants (Radloff, 1977). This scale consists of 20 items, and participants answer how often each item had occurred over the past week, with four response options ranging from 0, indicating

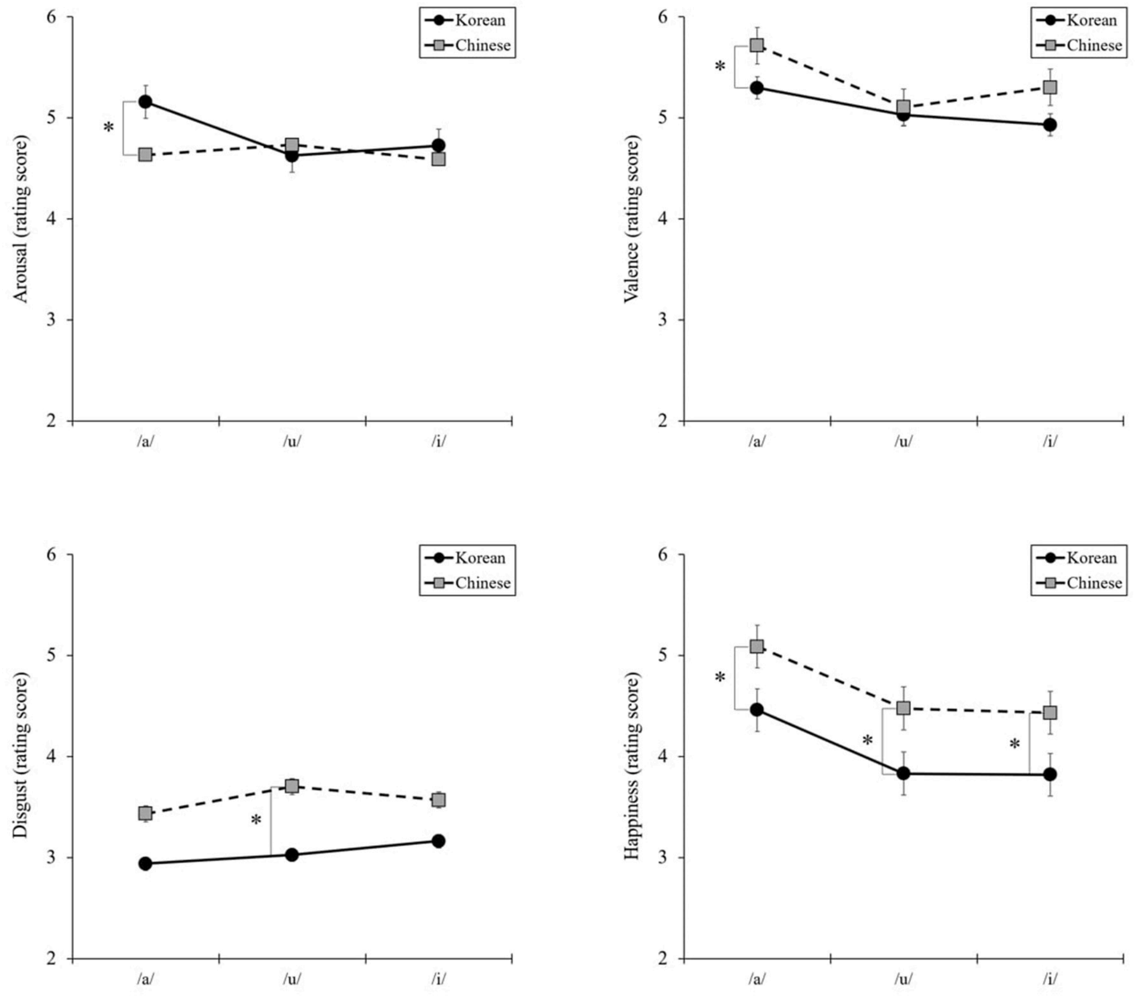


FIGURE 2
Interaction effects between three vowels and nationality in arousal, valence, and basic emotions: disgust and happiness. *$P < 0.05$.

"rarely," to 3, indicating "all times." Cronbach's alpha values of the K-CES-D and C-CES-D were 0.93 and 0.90, respectively.

Stimuli

For the experiment, 42 Hangul phonemic stimuli with artificial human sounds were created and used with a text-to-speech (TTS) program. All stimuli had an equal duration of 500 ms. Considering participants' fatigue, 42 voice stimuli were used combining three Korean vowels and 14 Korean consonants (Table 2). The three vowels used were those with the largest differences in pronunciation structure (Lee and Lee, 2000), and the 14 consonants used excluded double consonants. These consonants were classified into three types according to the articulation system of classification (Kim, 2019). Specifically, they were classified as lenis consonants if they did not require heavy breathing or straining of the throat, aspirated consonants if they involved the release of a burst of strong air during plosive sounds, and voiced consonants if they resonated in the mouth or nose when pronounced.

Data analysis

A dataset with 136 samples was included in the final analysis. As the first step in the analysis, a 2 (nationality: Korean, Chinese) × 3 (articulation: lenis, aspirated, voiced), 2 (nationality: Korean, Chinese) × 3 (vowel: /a/, /u/, /i/) two-way analysis of variance (ANOVA) was conducted to identify differences in sound symbolism between participants of different nationalities. In addition, a 2 (nationality: Korean, Chinese) × 2 (voice gender: female, male) two-way ANOVA was conducted to explore the arousal and valence patterns according to nationality and voice gender. A 2 (voice gender: female, male) × 3 (articulation: lenis, aspirated, voiced) and 2 (voice gender: female, male) × 3 (vowel: /a/, /u/, /i/) mixed ANOVA was conducted with Korean and Chinese participants, respectively, to explore the patterns of arousal and valence for each nationality. An independent sample t-test was performed for continuous variables to compare psychological characteristics and perform planned comparisons.

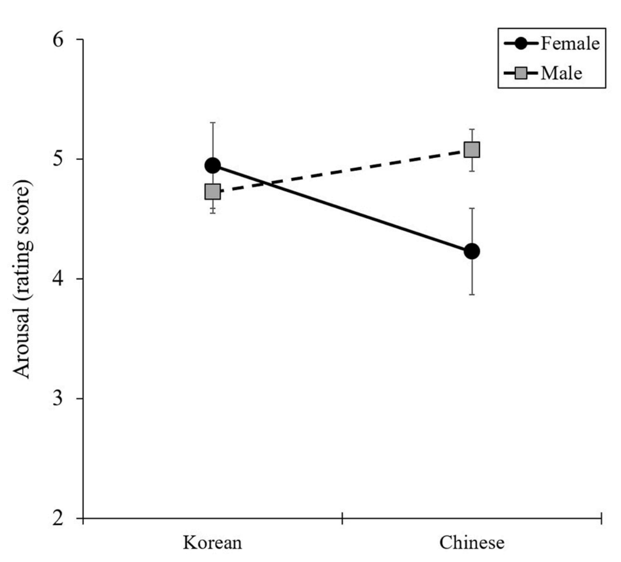


FIGURE 3
Interaction effects between nationality and voice gender in arousal.

Results

Differences between nationalities for articulation and vowels

Arousal, valence, and basic emotions were used as dependent variables, three types of articulation (lenis, aspirated, and voiced) and three vowels (/a/, /u/, and /i/) were used as within-subjects factors and nationality (Korean or Chinese) was used as a between-subjects factor to perform the repeated-measures ANOVA.

The results presented in Figure 1 indicate that the interaction between the three types of articulation and nationality was significant for arousal [$F_{(2,268)} = 11.93$, $P = 0.00$, $\eta^2 = 0.08$] and valence [$F_{(2,268)} = 7.33$, $P = 0.001$, $\eta^2 = 0.05$]. Then, planned comparisons were performed using independent sample t-tests, which revealed no differences between Korean and Chinese participants for lenis or voiced articulation. However, for aspirated articulation, arousal was higher in Chinese participants than in Korean participants [$t_{(134)} = 2.51$, $P = 0.01$, $d = 0.43$], whereas valence was higher in Korean than Chinese participants [$t_{(134)} = -3.09$, $P = 0.00$, $d = 0.53$]. Furthermore, significant interactions were observed between two basic emotions. At the lenis and voiced articulation levels, disgust was higher in Chinese participants than in Korean participants [$t_{(134)} = -1.99$, $P = 0.49$, $d = 0.34$; $t_{(134)} = -1.985$, $P = 0.049$, $d = 0.34$), and at the lenis and aspirated articulation levels, happiness was higher in Chinese participants than in Korean participants [$t_{(134)} = -2.61$, $P = 0.01$, $d = 0.45$; $t_{(134)} = -2.93$, $P = 0.004$, $d = 0.50$].

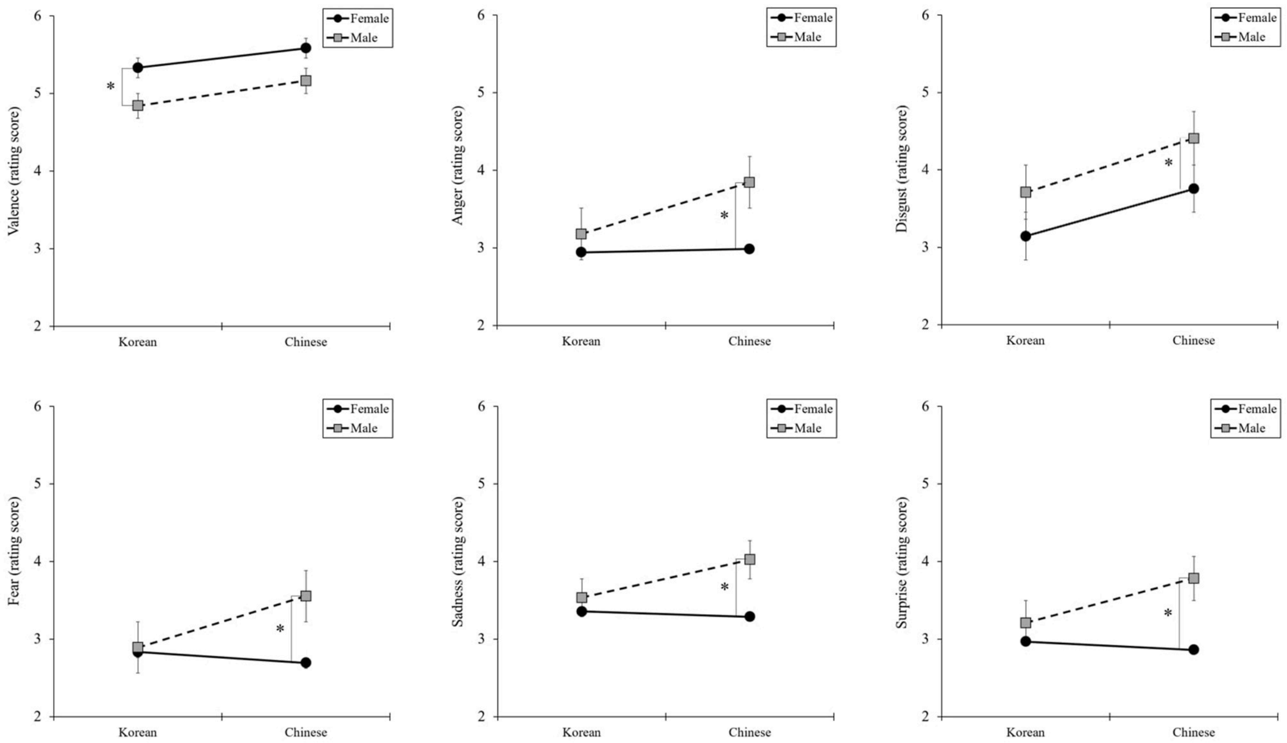


FIGURE 4
Simple main effect analysis of nationality and voice gender. *$P < 0.05$.

TABLE 3 Differences in the types of articulation for voice gender by nationality.

		Nationality	Voice gender		*F*
			Female	Male	
Arousal	Lenis	Korean	4.94 ± 0.66	4.71 ± 1.04	1.19
		Chinese	4.31 ± 1.56	4.94 ± 1.03	3.87
	Aspirated	Korean	5.68 ± 1.15	5.31 ± 1.40	1.38
		Chinese	4.52 ± 1.80	5.27 ± 1.08	4.31*
	Voiced	Korean	4.23 ± 1.01	4.15 ± 1.27	0.07
		Chinese	3.86 ± 1.57	5.03 ± 1.05	13.00**
Valence	Lenis	Korean	5.38 ± 0.67	4.81 ± 1.04	7.27**
		Chinese	5.54 ± 1.57	5.23 ± 0.72	1.11
	Aspirated	Korean	4.76 ± 0.92	5.14 ± 1.08	2.83
		Chinese	5.61 ± 1.52	5.45 ± 0.89	0.268
	Voiced	Korean	5.89 ± 1.06	4.59 ± 1.24	21.87***
		Chinese	5.60 ± 1.75	4.81 ± 0.76	5.83*
Anger	Lenis	Korean	2.89 ± 1.44	3.17 ± 1.58	0.60
		Chinese	3.09 ± 1.58	3.76 ± 1.55	3.08
	Aspirated	Korean	3.16 ± 1.63	3.29 ± 1.69	0.10
		Chinese	3.09 ± 1.60	3.93 ± 1.69	4.48*
	Voiced	Korean	2.77 ± 1.68	3.07 ± 1.66	0.56
		Chinese	2.78 ± 1.54	3.84 ± 1.53	8.21**
Disgust	Lenis	Korean	2.94 ± 1.55	3.17 ± 1.57	0.38
		Chinese	3.20 ± 1.61	3.98 ± 1.46	4.39*
	Aspirated	Korean	3.21 ± 1.67	2.98 ± 1.54	0.35
		Chinese	3.30 ± 1.78	3.83 ± 1.60	1.70
	Voiced	Korean	2.70 ± 1.70	3.25 ± 1.65	1.83
		Chinese	2.93 ± 1.61	4.18 ± 1.63	10.13**
Fear	Lenis	Korean	2.78 ± 1.55	2.91 ± 1.60	0.12
		Chinese	2.75 ± 1.60	3.53 ± 1.61	3.96
	Aspirated	Korean	3.08 ± 1.61	2.83 ± 1.57	0.42
		Chinese	2.74 ± 1.62	3.53 ± 1.70	3.79
	Voiced	Korean	2.65 ± 1.77	2.94 ± 1.67	0.51
		Chinese	2.58 ± 1.70	3.61 ± 1.62	6.51*
Sadness	Lenis	Korean	3.26 ± 1.44	3.65 ± 1.69	1.06
		Chinese	3.32 ± 1.71	4.03 ± 1.41	3.47
	Aspirated	Korean	3.38 ± 1.64	2.98 ± 1.30	1.22
		Chinese	3.21 ± 1.66	3.61 ± 1.42	1.12
	Voiced	Korean	3.43 ± 1.62	3.96 ± 1.87	1.59
		Chinese	3.33 ± 1.54	4.44 ± 1.35	10.03**
Surprise	Lenis	Korean	2.90 ± 1.54	3.17 ± 1.60	0.50
		Chinese	2.94 ± 1.62	3.74 ± 1.51	4.49*
	Aspirated	Korean	3.20 ± 1.66	3.50 ± 1.78	0.52
		Chinese	2.98 ± 1.59	3.90 ± 1.56	5.72*
	Voiced	Korean	2.81 ± 1.67	2.96 ± 1.54	0.16
		Chinese	2.66 ± 1.49	3.71 ± 1.55	8.06**

(Continued)

TABLE 3 (Continued)

		Nationality	Voice gender		*F*
			Female	Male	
Happiness	Lenis	Korean	3.93 ± 1.31	3.96 ± 1.50	0.01
		Chinese	4.60 ± 1.61	4.57 ± 1.33	0.01
	Aspirated	Korean	3.59 ± 1.41	4.53 ± 1.679	6.155*
		Chinese	4.71 ± 1.65	4.98 ± 1.373	0.507
	Voiced	Korean	4.44 ± 1.58	3.78 ± 1.664	2.851
		Chinese	4.91 ± 1.63	4.22 ± 1.411	3.505

*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$.

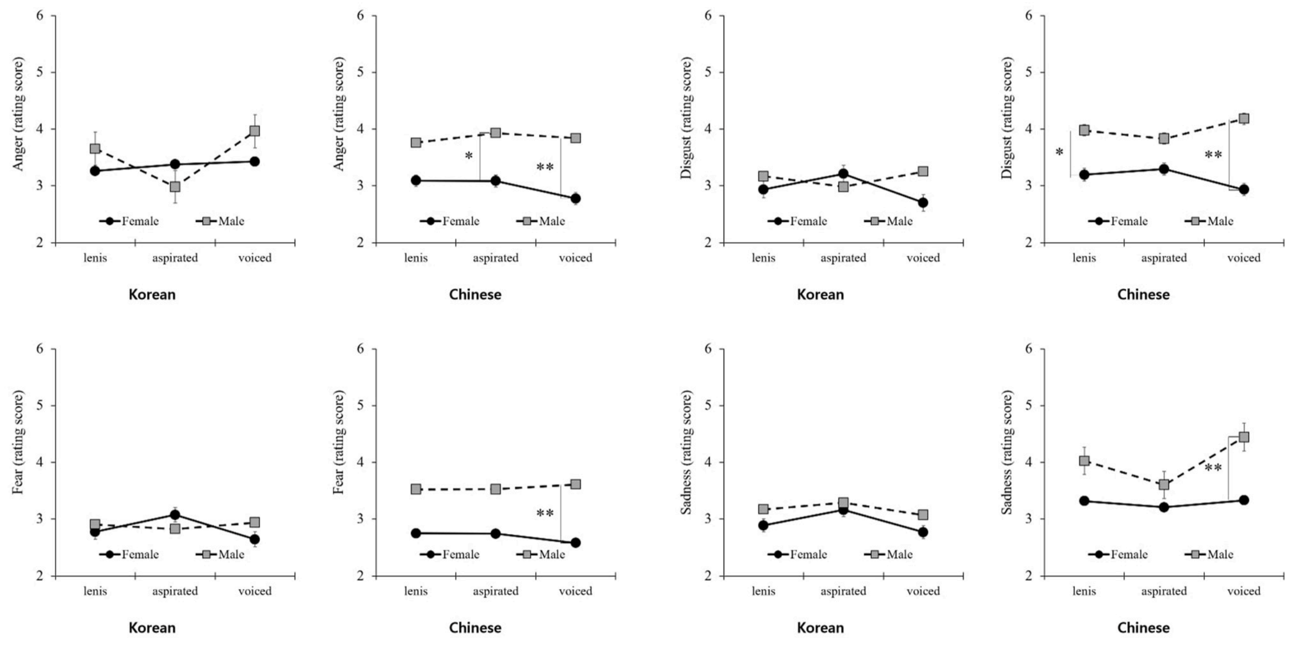


FIGURE 5
Differences in basic emotions for voice gender by nationality. *$P < 0.05$, **$P < 0.01$.

Similar results were found for vowels. As shown in Figure 2, the interaction between three vowels and nationality was significant in arousal [$F_{(2,268)} = 8.73$, $P = 0.000$, $\eta^2 = 0.06$] and valence [$F_{(2,268)} = 3.31$, $P = 0.04$, $\eta^2 = 0.02$]. Planned comparisons performed using independent sample t-tests showing no differences were found between Korean and Chinese for /u/ or /i/. However, for /a/, arousal was higher in Korean participants than in Chinese participants [$t_{(134)} = 2.40$, $P = 0.02$, $d = 0.41$], whereas the valence was higher in Chinese participants than in Korean participants [$t_{(134)} = -2.08$, $P = 0.04$, $d = 0.36$]. Furthermore, significant interactions were observed between the two basic emotions. For /u/, disgust was higher in Chinese participants than in Korean participants [$t_{(134)} = -2.40$, $P = 0.02$, $d = 0.41$], and for all vowels, happiness was higher in Chinese participants than in Korean participants [/a/: $t_{(134)} = -2.32$, $P = 0.02$, $d = 0.40$; /u/: $t_{(134)} = -2.59$, $P = 0.01$, $d = 0.44$; /i/: $t_{(134)} = -2.35$, $P = 0.02$, $d = 0.40$]. However, no significant interaction effects were found for the other basic emotions.

Differences between nationalities for articulation and vowels

The results of the two-way ANOVA, presented in Figure 3, provide statistical support for the interaction between nationality (Korean or Chinese) and voice gender (female or male) for arousal [$F_{(1,132)} = 7.92$, $P = 0.006$, $\eta^2 = 0.06$]. However, no significant interaction effects were found for valence or basic emotions.

No interaction effects were found on the valence or basic emotional score. However, the one-way ANOVA for simple main-effect analysis revealed differences in voice gender by nationality (Figure 4). Unlike Korean participants, Chinese participants reported feeling more negatively toward male voices than female voices.

TABLE 4 Differences in the value of vowels for voice gender by nationality.

		Nationality	Voice gender		F
			Female	Male	
Arousal	/a/	Korean	5.13 ± 0.82	5.19 ± 1.18	0.07
		Chinese	4.14 ± 1.66	5.13 ± 1.10	8.24**
	/u/	Korean	4.64 ± 0.72	4.61 ± 1.33	0.01
		Chinese	4.36 ± 1.66	5.10 ± 0.99	5.0*
	/i/	Korean	5.10 ± 0.85	4.37 ± 1.12	8.59**
		Chinese	4.18 ± 1.67	5.00 ± 1.00	6.13*
Valence	/a/	Korean	5.48 ± 0.74	5.12 ± 1.09	2.49
		Chinese	5.88 ± 1.72	5.56 ± 0.86	0.94
	/u/	Korean	5.29 ± 0.85	4.78 ± 1.14	4.35*
		Chinese	5.22 ± 1.52	4.99 ± 0.85	0.59
	/i/	Korean	5.23 ± 0.74	4.63 ± 1.12	6.84*
		Chinese	5.66 ± 1.62	4.95 ± 0.94	4.92*
Anger	/a/	Korean	2.93 ± 1.58	3.23 ± 1.74	0.55
		Chinese	2.90 ± 1.43	3.74 ± 1.53	5.54*
	/u/	Korean	2.91 ± 1.55	3.13 ± 1.70	0.31
		Chinese	3.13 ± 1.72	3.97 ± 1.57	4.45*
	/i/	Korean	2.98 ± 1.53	3.18 ± 1.66	0.25
		Chinese	2.93 ± 1.53	3.80 ± 1.64	5.33*
Disgust	/a/	Korean	2.84 ± 1.59	3.04 ± 1.64	0.25
		Chinese	3.06 ± 1.59	3.80 ± 1.66	3.53
	/u/	Korean	2.96 ± 1.65	3.09 ± 1.58	0.11
		Chinese	3.31 ± 1.83	4.10 ± 1.46	3.85
	/i/	Korean	3.05 ± 1.52	3.28 ± 1.66	0.34
		Chinese	3.05 ± 1.62	4.09 ± 1.52	7.36**
Fear	/a/	Korean	2.69 ± 1.64	2.86 ± 1.58	0.17
		Chinese	2.66 ± 1.64	3.53 ± 1.66	4.75*
	/u/	Korean	2.83 ± 1.61	2.85 ± 1.58	0.00
		Chinese	2.76 ± 1.68	3.56 ± 1.53	4.18*
	/i/	Korean	2.98 ± 1.63	2.98 ± 169	0.00
		Chinese	2.66 ± 1.58	3.58 ± 1.72	5.24*
Sadness	/a/	Korean	3.16 ± 1.56	3.12 ± 1.48	0.01
		Chinese	3.13 ± 1.60	3.72 ± 1.46	2.54
	/u/	Korean	3.62 ± 1.65	3.70 ± 1.62	0.04
		Chinese	3.48 ± 1.75	4.01 ± 1.29	2.01
	/i/	Korean	3.29 ± 1.55	3.78 ± 1.75	1.52
		Chinese	3.24 ± 1.57	4.34 ± 1.51	8.59**
Surprise	/a/	Korean	2.94 ± 1.61	3.58 ± 1.78	2.45
		Chinese	3.01 ± 1.47	3.83 ± 1.45	5.37*
	/u/	Korean	2.99 ± 1.67	3.11 ± 1.65	0.09
		Chinese	2.80 ± 1.63	3.84 ± 1.61	7.04*
	/i/	Korean	2.98 ± 1.54	2.94 ± 1.52	0.01
		Chinese	2.78 ± 1.57	3.68 ± 1.65	5.35*

(Continued)

TABLE 4 (Continued)

		Nationality	Voice gender		*F*
			Female	Male	
Happiness	/a/	Korean	4.36 ± 1.95	4.56 ± 1.76	0.26
		Chinese	5.00 ± 1.74	5.17 ± 1.42	0.20
	/u/	Korean	3.72 ± 1.46	3.95 ± 1.46	0.43
		Chinese	4.53 ± 1.59	4.43 ± 1.32	0.07
	/i/	Korean	3.88 ± 1.34	3.76 ± 1.63	0.12
		Chinese	4.70 ± 1.64	4.17 ± 1.42	2.04

$^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

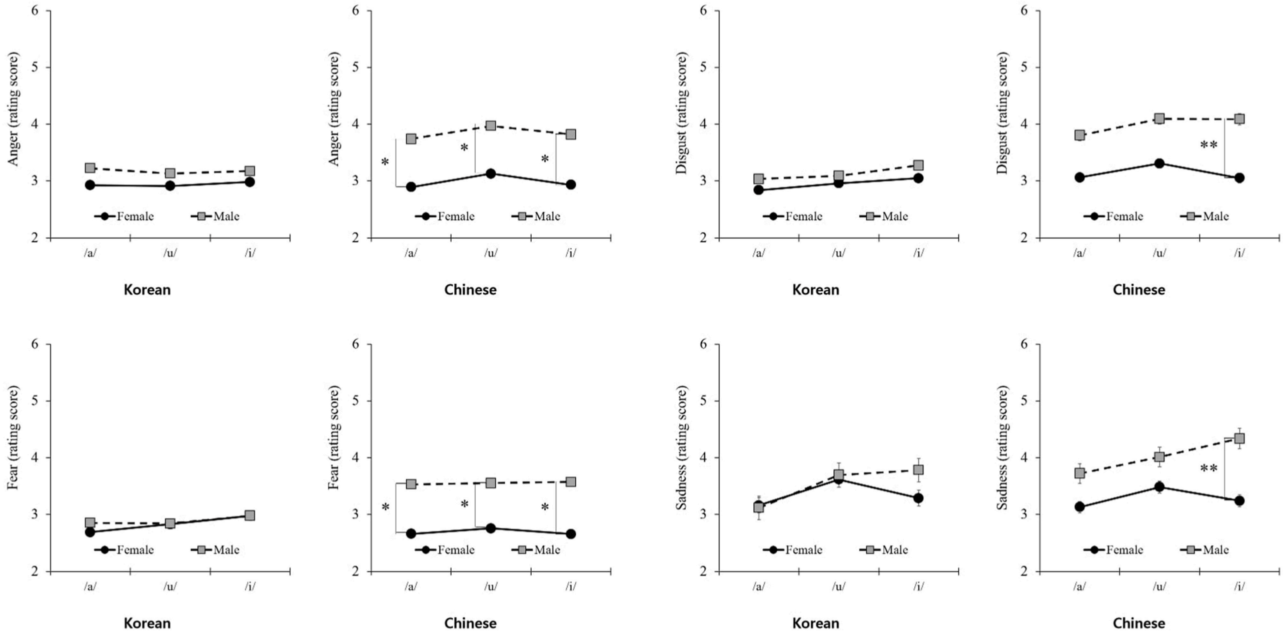


FIGURE 6
Differences in basic emotions for voice gender by nationality. $^*P < 0.05$, $^{**}P < 0.01$.

Differences in the value of articulation for voice gender by nationality

As shown in Table 3, the articulation types showed differences in arousal based on voice gender by nationality. Aspirated articulation elicited more arousal than voiced articulation, regardless of voice gender. However, unlike Korean participants, Chinese participants showed significant differences in arousal based on voice gender. In particular, for aspirated and voiced articulation, articulation ratings were significantly higher for a male voice than for a female voice. However, valence exhibited a different pattern for the male voice compared to the female voice. Although arousal was rated lower for the female voice than for the male voice, valence was rated higher for the female voice than for the male voice. Regarding basic emotions, clear differences were observed in the patterns between nationalities (Figure 5). Korean participants showed no significant differences in scores regardless of voice gender, whereas Chinese participants exhibited a difference in scores between voice and gender, especially for voiced articulation.

Differences in the value of vowels for voice gender by nationality

As presented in Table 4, differences were found in arousal for the three vowels based on voice gender by nationality. Unlike the articulation results, for vowels, the patterns of arousal for voice gender differed depending on nationality. In Chinese participants, the difference in arousal value based on voice gender was remarkable for all types; however, in Korean participants, a difference in arousal value based on voice gender was found only for /i/. Regarding basic emotions, clear differences were observed in the patterns between nationalities. For Korean participants, the scores were not significantly different, regardless of voice gender. However, for Chinese participants, scores differed between voice and gender (Figure 6).

Discussion and conclusion

This study aimed to explore cultural differences by comparing the degree of arousal and valence experienced by Korean and Chinese women in Hangul phonemes, based on the gender of an AI voice. The results of this study revealed significant differences in arousal levels between Korean and Chinese women in response to male AI voices. In particular, Chinese women exhibited distinct differences in emotional perceptions of male and female voices in voiced consonants. In addition, this study classified participants by nationality and identified cultural differences in arousal and valence patterns according to articulation and vowels.

This study revealed that arousal and valence levels differed between Korean and Chinese women, even for phonemic units without conceptual meaning. This is consistent with Russell's claim that emotional stimuli may have cultural differences. For vowels, the results contradict those of previous studies that suggest universal emotional responses, depending on the culture. This disparity is likely because China uses a tonal language, unlike Korea. While Korean has a dialect with a different pitch from that of the standard language, it is not a tone language that conveys variations in meaning through pitch differences. By contrast, China's tonal language facilitates meaning changes through changes in tone. Therefore, Korean and Chinese listeners may experience differences in arousal and valence patterns when hearing the same sound. This finding is supported by the results of a study comparing the vowels /a/, /u/, and /i/ in Chinese with tone and English without tone, highlighting differences in sound symbolism based on the presence or absence of lexical tones (Chang et al., 2021).

We aimed to observe the differences between wakefulness and emotional outcomes. Clear and distinct differences were apparent in wakefulness, whereas in emotion, only the interaction between nationality and consonants proved significant. To elucidate, the most substantial difference between Korean and Chinese syllables, apart from phonemic constraints, lies in the pronunciation elements through the presence of sound. Korean allows for seven syllable-final consonants: ㅂ /p/, ㄷ /t/, ㄱ /k/, ㅁ /m/, ㄴ /n/, ㅇ /η, and ㄹ /l. By contrast, in Chinese, syllable-final consonants are restricted to two: /-n/ and /η/. The disparity in syllable-consonant combination constraints is the primary cause of phonological variation between Korean and Chinese.

The importance lies not only in the difference in the phoneme itself but also in the interaction effect between nationality and voice gender. Differences in the arousal and valence experienced in response to Hangul phonemes varied by nationality depending on whether the voice was female or male. In particular, Chinese women were found to experience negative emotions even when the voiced sound was presented with a male voice, although voiced consonant is an articulation method that results in less arousal than lenis and aspirated consonants. Although the study was conducted among Chinese participants living in Korea rather than in China, cultural values do not change easily (Hofstede, 1984, 1998), which may be attributed to cultural differences based on gender roles. According to a cultural-level study, China has a wider gender power gap than Korea; China highly values the image of masculinity (Moon and Woo, 2019), and Chinese women feel that they are not free to express themselves and are restricted in their opportunities to demonstrate their abilities because of men (Sun, 2022). Consequently, Chinese women are less dependent on men and experience a sense of competition with them. Therefore, compared with Korean participants, Chinese participants experienced more negative emotions toward a male voice than a female voice. This trend can be further clarified through research on cultural differences in gender-related issues.

In recent years, the generation and interpretation of literary works and various forms of literature through AI have highlighted the increasing importance of studying people's emotional responses. Specifically, the examination of the ability of AI to convey emotions or impart specific emotions is important. Currently, by utilizing natural language processing and machine learning techniques, research studies investigate how emotions are expressed and recognized in literary works, as well as how the tone and expression AI employs when narrating stories evoke emotional responses in listeners (Spezialetti et al., 2020; Lettieri et al., 2023). This study aimed to determine the emotional responses elicited in listeners by the tones and expression styles used by AI when delivering verbal expressions. Furthermore, in an era marked by an active international approach to media use, this study is significant in exploring the specific differences in emotional responses to voices in Chinese and Korean, languages characterized by distinct intonations, and cultural and political expressions despite their geographical proximity.

Limitations

This study has several limitations. First, although the study focused on AI, various AI voices were not used for the investigation. Various AI speakers have been released in both Korea and China, and people exhibit different preferences. Therefore, the sound stimuli used in this study are not perfectly consistent with the degree of arousal and valence experienced with the currently available voices of the AI speakers. For marketing applications, further research using the voices of various AI speakers is needed. In addition, some AI speakers now allow customers to purchase celebrity voices that they like directly, in which case, the results of this study would be challenging to apply, even for male voices.

Second, the study did not include comparisons with other cultures. Both Korea and China are cultural regions in Northeast Asia, but subtle differences may occur across various languages and cultures. Although Korea and China are in the same region, the fact that differences were found according to voice gender indicates that differences based on voice gender may occur in other cultures and may be relatively larger. In the future, multicultural studies should be conducted to compare a wide range of languages and cultures.

Third, this study was conducted with female participants. Women tend to respond more emotionally than men, and only women were recruited based on the existing claim that sound symbolism does not differ significantly by gender. However, cultural differences may interact with gender.

Finally, this study used only three vowels (/a/, /u/, and /i/) and 14 consonants; however, fortis consonants were not used. In Chinese, intonation plays a significant role alongside pronunciation in conveying the meanings of individual words. Hence, future research should meticulously examine variations in intonation within words of identical pronunciation to ascertain their emotional impact conveyed through speech. In addition to articulation, intonation is also used in China. Therefore, a more detailed classification is required to reflect the actual situation.

Despite these limitations, this study is significant in demonstrating that arousal and valence may differ in articulation types and vowels depending on cultural differences, and that voice gender can also affect perceived emotions. This principle supports sound symbolism and has practical implications for voice gender and branding in AI applications.

Data availability statement

All data can be made available upon request to the corresponding author.

Ethics statement

The studies involving humans were approved by the Institutional Review Board of Chung-Ang University (IRB NO. 1041078-20230210-HR-032). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

M-SL: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing— original draft. G-EL: Conceptualization, Data curation, Formal analysis, Project administration, Validation, Writing—original draft, Writing—review & editing, Supervision. SL: Conceptualization, Funding acquisition, Project administration, Validation, Writing—review & editing, Supervision. J-HL: Conceptualization, Project administration, Validation, Writing—review & editing, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2022S1A5C2A04093121).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aveyard, M. E. (2012). Some consonants sound curvy: effects of sound symbolism on object recognition. *Mem. Cogn.* 40, 83–92. doi: 10.3758/s13421-011-0139-3

Barrett, L. F. (2006a). Are emotions natural kinds?. *Perspect. Psychol. Sci.* 1, 28–58. doi: 10.1111/j.1745-6916.2006.00003.x

Barrett, L. F. (2006b). Valence is a basic building block of emotional life. *J. Res. Pers.* 40, 35–55. doi: 10.1016/j.jrp.2005.08.006

Chang, Y., Zhao, M., Chen, Y., and Huang, P. (2021). The effects of mandarin chinese lexical tones in sound–shape and sound–size correspondences. *Multisens. Res.* 35, 243–257. doi: 10.1163/22134808-bja10068

Chen, Y. C., Huang, P. C., Woods, A., and Spence, C. (2016). When "Bouba" equals "Kiki": cultural commonalities and cultural differences in sound-shape correspondences. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep26681

Chi, I., and Boey, K. W. (1993). Hong Kong validation of measuring instruments of mental health status of the elderly. *Clin. Gerontol.* 13, 35–51. doi: 10.1300/J018v13n04_04

Chon, K. K., Choi, S. C., and Yang, B. C. (2001). Integrated adaptation of CES-D in Korea. *Kor. J. Health Psychol.* 6, 59–76.

Ekman, P., Levenson, R. W., and Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science* 221, 1208–1210. doi: 10.1126/science.6612338

Ekman, P., and Oster, H. (1979). Facial expressions of emotion. *Annu. Rev. Psychol.* 30, 527–554. doi: 10.1146/annurev.ps.30.020179.002523

Ernst, C. P., and Herm-Stapelberg, N. (2020). "Gender Stereotyping's Influence on the Perceived Competence of Siri and Co.," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 4448–4453. doi: 10.24251/HICSS.2020.544

Grühn, D., and Scheibe, S. (2008). Age-related differences in valence and arousal ratings of pictures from the International Affective Picture System (IAPS): Do ratings become more extreme with age? *Behav. Res. Methods* 40, 512–521. doi: 10.3758/BRM.40.2.512

Hahn, D. W., and Kang, H. J. (2000). Appropriateness and frequency of emotion terms in Korea. *Kor. J. Psychol. General* 19, 78–98.

Hofstede, G. (1984). *Culture's Consequences: International Differences in Work-Related Values, 2nd Edn.* Beverly Hills, CA: SAGE Publications.

Hofstede, G. (1998). Identifying organizational subcultures: an empirical approach. *J. Manag. Studi.* 35, 1–12. doi: 10.1111/1467-6486.00081

Hoshi, H., Kwon, N., Akita, K., and Auracher, J. (2019). Semantic associations dominate over perceptual associations in vowel–size iconicity. *Iperception* 10, 1–31. doi: 10.1177/2041669519861981

Huang, L., Yang, T., and Li, Z. (2003). Applicability of the positive and negative affect scale in Chinese. *Chin. Mental Health J.* 17, 54–56.

Jang, J. H., and Ju, D. Y. (2019). "Usability test of emotional speech from AI speaker," in *Proceedings of The HCI Society of Korea*, 705–712.

Ji, W., Liu, R., and Lee, S. (2019). "Do drivers prefer female voice for guidance? An interaction design about information type and speaker gender for autonomous driving car," in *International Conference on Human-Computer Interaction,* 208–224. doi: 10.1007/978-3-030-22666-4_15

Kim, J. T., and Shin, D. K. (1978). A study of based on the standardization of the STAI for Korea. *New Med. J.* 21, 69–75.

Kim, N. Y., and Yun, J. Y. (2021). User experience research on sex and pitch of AI agent's voice based on the purpose and context of the utterance. *Design Converg. Study* 20, 109–130. doi: 10.31678/SDC89.8

Kim, S. H. (2019). The characteristics of consonantal distribution in Korean sound-symbolic words. *Stud. Phonet. Phonol. Morphol.* 25, 387–414. doi: 10.17959/sppm.2019.25.3.387

Lee, H. H., Kim, E. J., and Lee, M. K. (2003). A validation study of Korea positive and negative affect schedule: The PANAS scales. *Kor. J. Clin. Psychol.* 22, 965–946.

Lee, J. Y., and Lee, S. H. (2000). A study of consonant perception and production by children with profound sensorineural hearing loss. *Commun. Sci. Disord.* 5, 1–17.

Lettieri, G., Handjaras, G., Bucci, E., Pietrini, P., and Cecchetti, L. (2023). How male and female literary authors write about affect across cultures and over historical periods. *Affect. Sci.* 4, 770–780. doi: 10.1007/s42761-023-00219-9

Libkuman, T. M., Otani, H., Kern, R., Viger, S. G., and Novak, N. (2007). Multidimensional normative ratings for the international affective picture system. *Behav. Res. Methods* 39, 326–334. doi: 10.3758/BF03193164

Lowrey, T. M., Shrum, L. J., and Dubitsky, T. M. (2003). The relation between brand-name linguistic characteristics and brand-name memory. *J. Advert.* 32, 7–17. doi: 10.1080/00913367.2003.10639137

Maurer, D., Pathman, T., and Mondloch, C. J. (2006). The shape of boubas: Sound–shape correspondences in toddlers and adults. *Dev. Sci.* 9, 316–322. doi: 10.1111/j.1467-7687.2006.00495.x

Moon, S.-J., and Woo, B. (2019). The cultural values and the interpersonal communication of the young chinese immigrants residing in Korea. *J. Multicult. Soc.* 12, 79–102. doi: 10.14431/jms.2019.02.12.1.79

Obinali, C. (2019). "The perception of gender in voice assistants," in *Proceedings of the Southern Association for Information Systems Conference.*

Park, E. J., Kikutani, M., Yogo, M., Suzuki, N., and Lee, J. H. (2018). Influence of culture on categorical structure of emotional words: Comparison between Japanese and Korean. *J. Cross Cult. Psychol.* 49, 1340–1357. doi: 10.1177/0022022118789789

Pejovic, J., and Molnar, M. (2017). The development of spontaneous sound-shape matching in monolingual and bilingual infants during the first year. *Dev. Psychol.* 53:581. doi: 10.1037/dev0000237

Poon-Feng, K., Huang, D. Y., Dong, M., and Li, H. (2014). "Acoustic emotion recognition based on fusion of multiple feature-dependent deep Boltzmann machines," in *The 9th International Symposium on Chinese Spoken Language Processing* (IEEE), 584–588. doi: 10.1109/ISCSLP.2014.6936696

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Appl. Psychol. Meas.* 1, 385–401. doi: 10.1177/014662167700100306

Ramachandran, V. S., and Hubbard, E. M. (2001). Synaesthesia–a window into perception, thought and language. *J. Conscious. Stud.* 8, 3–34.

Rogers, S. K., and Ross, A. S. (1975). A cross-cultural test of the Maluma-Takete phenomenon. *Perception* 4, 105–106. doi: 10.1068/p040105

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178. doi: 10.1037/h0077714

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychol. Rev.* 110, 145–172. doi: 10.1037/0033-295X.110.1.145

Sapir, E. (1929). A study in phonetic symbolism. *J. Exp. Psychol.* 12:225. doi: 10.1037/h0070931

Shang, N., and Styles, S. J. (2017). Is a high tone pointy? Speakers of different languages match Mandarin Chinese tones to visual shapes differently. *Front. Psychol.* 8:2139. doi: 10.3389/fpsyg.2017.02139

Shinohara, K., and Kawahara, S. (2010). A cross-linguistic study of sound symbolism: the images of size. *Ann. Meet. Berk. Linguist. Soc.* 36, 396–410. doi: 10.3765/bls.v36i1.3926

Slunecko, T., and Hengl, S. (2006). Culture and media: a dynamic constitution. *Cult. Psychol.* 12, 69–85. doi: 10.1177/1354067X06061594

Spezialetti, M., Placidi, G., and Rossi, S. (2020). Emotion recognition for human-robot interaction: recent advances and future perspectives. *Front. Robot. AI* 7:532279. doi: 10.3389/frobt.2020.532279

Spielberger, C. D., Gorsuch, R. L., and Lushene, R. E. (1970). *STAI Manual for a State-Trait Anxiety Inventory*. Consulting Psychologist Press.

Styles, S. J., and Gawne, L. (2017). When does Maluma/Takete fail? Two key failures and a meta-analysis suggest that phonology and phonotactics matter. *i-Perception* 8:2041669517724807. doi: 10.1177/2041669517724807

Sun, Y. (2022). "Gender conflict in China in the context of new media," in *Proceedings of the 2021 International Conference on Social Development and Media Communication (SDMC 2021),* 1385–1389. doi: 10.2991/assehr.k.220105.253

Tsoi, M. M., Ho, E., and Mak, K. C. (1986). "Becoming pregnant again after stillbirth or the birth of a handicapped child," in *Hormone and Behavior*, eds. L. Dennerstein and I. Fraser (Holland, MI: Elsevier Science), 310–316.

Val-Calvo, M., Álvarez-Sánchez, J. R., Ferrández-Vicente, J. M., and Fernández, E. (2020). Affective robot story-telling human-robot interaction: exploratory real-time emotion estimation analysis using facial expressions and physiological signals. *IEEE Access* 8, 134051–134066. doi: 10.1109/ACCESS.2020.3007109

Wang, C., Teo, T. S., and Janssen, M. (2021). Public and private value creation using artificial intelligence: an empirical study of AI voice robot users in Chinese public sector. *Int. J. Inf. Manage.* 61:102401. doi: 10.1016/j.ijinfomgt.2021.102401

Yu, S. P. (2021). *The Gleam-Glum Effect with Pseudo-Words:/i/vs/Λ/Phonemes Carry Emotional Valence that Influences Semantic Interpretation* (Doctoral Dissertation). Arizona State University. doi: 10.1037/xlm0001017

Zheng, W. Q., Yu, J. S., and Zou, Y. X. (2015). "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (IEEE), 827–831. doi: 10.1109/ACII.2015.7344669

Frontiers in Psychology

TYPE Original Research
PUBLISHED 07 August 2024
DOI 10.3389/fpsyg.2024.1409646

A research on copyright issues impacting artists emotional states in the framework of artificial intelligence

Hüseyin Kambur[1†] and Ayhan Dolunay[2*†]

[1]Faculty of Communication, Near East University, Nicosia, Cyprus, [2]Faculty of Communication, Grand Library, Near East University, Nicosia, Cyprus

OPEN ACCESS

EDITED BY
Simone Belli,
Complutense University of Madrid, Spain

REVIEWED BY
Zhi Yang,
Griffith University, Australia
Evgenia Gkintoni,
General University Hospital of Patras, Greece

*CORRESPONDENCE
Ayhan Dolunay
✉ ayhan.dolunay@neu.edu.tr

†These authors have contributed equally to this work

RECEIVED 30 March 2024
ACCEPTED 23 July 2024
PUBLISHED 07 August 2024

CITATION
Kambur H and Dolunay A (2024) A research on copyright issues impacting artists emotional states in the framework of artificial intelligence.
Front. Psychol. 15:1409646.
doi: 10.3389/fpsyg.2024.1409646

Art and artistic creation serve as a means for artists to communicate with their environment, society, and the external world. However, the protection of artistic creations, as forms of communication, is not only a right for artists but also serves as a crucial safeguard that nurtures them during the creative process. Beyond the traditional issues of copyright, the significant advancements in Artificial Intelligence (AI) in today's digital world have introduced a new debate regarding the ownership of copyright in artistic creations generated by AI. The question arises whether copyright belongs to the AI itself or to the individuals who guide the creative process behind it. In this study, based on the concepts of art, artistic creation, and emotional states, copyright issues will be examined. Data obtained from semi-structured in-depth interviews with artists and academic experts (eight artists, two communication experts, two law experts, and eight psychology experts) in the field will be analysed through content analysis to explore their perspectives regarding the discussion on emotional states, AI, and copyrights. The research highlights the variability of emotional states and their significant effects on individuals. Addressing the increasing trend of copyright issues, particularly within the framework of digitalization and inadequate legal regulations, it was found that artists' emotional states are negatively impacted by these problems. This negative influence can adversely affect artists' creativity and desire to produce. On the other hand, it was also identified that in artworks produced especially through AI, if artists' rights are not protected, there is a possibility of negative emotional states arising. In conclusion, suggestions are as follows: Emphasising the importance of awareness-raising educational activities nationally and internationally, national copyright law (in Northern Cyprus) needs to be revised to protect traditional copyright and be expanded to include digital copyright, especially for works produced through AI. On an international level, emphasising the need to revise international agreements to include regulations for works produced through AI or to create a new agreement based on the importance of this issue.

KEYWORDS

copyright, artificial intelligence, art, artworks, artists, emotional states

1 Introduction

Exploring the question of what art is has been a challenging process throughout history, with philosophers seeking answers from ancient times to the present day. The quest for an artistic response to existential theories constitutes the fundamental reason for this pursuit. Plato, particularly, sought to answer the question of art by assuming a speculative approach, considering it as a reflection of entities (Barasch, 2013). Aristotle, on the other hand, presented a similar notion in a different manner, defining art as imitation (Erinç, 1988).

Regardless of the theoretical explanation of art, artworks and artists should be accepted as a whole. Copyright laws are crucial in protecting both the artwork and the artist. Idealist thinkers argue that artworks should serve artistic thought rather than aiming to make money. This is why they may not even consider populist art genres as genuine art (Balkır, 2020).

While many scholars and thinkers agree that this mindset may hinder the development of art, they often remain silent when it comes to making a profit through art (Haiven, 2015). On the other hand, despite the emotional importance of art in human life, evaluating it in a material sense is challenging. Art holds significant importance as a means of emotional expression and communication for humanity, and it is also of critical importance for artists. This is because creating artworks serves as both a means of nourishing their souls, so to speak, and as a fundamental mechanism for sustaining their livelihoods. At this point, it may be difficult to assess art's emotional significance in a monetary context, it is natural for artists to pursue their profession not only for the sake of art but also to earn a living (Balkır, 2020).

At this point, obtaining copyright for their ideas and/or artworks is crucial for artists to be able to generate income (Akdoğan, 2001). Copyright is extremely important for artists who produce artworks. Copyright is necessary for artists to protect their rights, ideas, and earnings both nationally and internationally. In order for an artist to claim rights over their own work, they must own the copyright to the work, enabling them to prevent intellectual and artistic theft in this way (Akipek and Dardağan, 2001).

Within this framework, copyright ownership on artworks is considered highly important for artists. In this context, it can be argued that copyright infringements may lead to negative emotional states in artists, and these negative emotional states may negatively impact creativity and productivity like a domino effect. While this situation holds true even in traditional copyright approaches, in today's digital age, the ownership of copyright for artworks produced by AI is a significant topic of debate.

In this regard, Gillotte (2020) has addressed the issue of copyright ownership in artworks created with AI in today's digital age with the following statement: "Assuming that an AI-generated work is copyrightable, we turn to the question of who owns that copyright." Additionally, the US Copyright Office and many academics have argued that copyright cannot belong to AI due to its lack of legal personality (Gillotte, 2020). However, some authors also suggest that copyright could be shared between the artist who contributed to the creation of the work and the AI (Darvishi et al., 2022).

At this point, the central question or problem of this issue revolves around whether the copyright of relevant works belongs to AI, the individual guiding it, or both. Hence, this issue needs to be carefully addressed. Because, the emotional states of artists significantly influence their productivity and creativity (Flaherty, 2011). An approach that may be considered as a violation of rights by artists could lead to negative emotional states and the aforementioned effects for their creativity and productivity too.

2 Basic concepts

2.1 Artificial intelligence

Initially introduced in academic articles, the concept of AI (See McCulloch and Pitts, 1943; Turing, 1950) gained prominence with the Dartmouth Conference (McCarthy et al., 2006). The ability of AI to solve intricate problems was acknowledged through the programme "Logic Theory" (Newell and Simon, 1956).

In the subsequent years, criticism arose regarding the limited capabilities of AI, demonstrating that it was not yet on par with human intelligence (See Dreyfus, 1972). However, with advancing and evolving technology, numerous studies addressed the progressing abilities of AI and its transformation (LeCun et al., 2015; Russell and Norvig, 2022).

The concept of AI can be defined as a system with abilities similar to human intelligence, fundamentally capable of performing tasks related to computer structures (Russell and Norvig, 2022). Comprising an interdisciplinary whole, AI possesses critical skills such as deep problem-solving, learning, and decision-making (Nilsson, 2010; Dolunay, 2024).

AI, which can generally be categorised into weak and strong AI, exhibits capabilities comparable to human intelligence in certain specified aspects for weak AI, whereas strong AI is closer to general intelligence levels (Kurzweil, 2005).

Having a vast application scope, AI manifests its impact in various fields such as the functioning of automatic systems, content analysis, healthcare, education, communication, etc. (Bengio, 2021). The examination of AI's usage in the mentioned fields is indeed important; however, the focus of this study will be on its application in artistic production within the given context. Prior to delving into this subject, it would be pertinent to briefly define the concepts of art, artist, and artwork.

2.2 Art, artist and artwork

Art can be defined as a process resulting from the combination of individuals' productivity, imagination, and aesthetic perception, culminating in a unified whole and subsequently expressed outwardly. The concept of art, with its wide boundaries, not only provides aesthetic and beauty satisfaction but also serves as a powerful tool for social critique, cultural expression, and the transmission of personal emotions and thoughts (Turgut, 1991). The broadness of the concept of art is due to its many subfields. Art encompasses numerous branches such as painting, music, theatre, sculpture, dance, literature, etc. (Adajian, 2022).

According to Adorno (1997), art possesses an autonomous structure, leading to profound effects both individually and socially. In this context, as stated, art shapes the identities, histories, and cultural values of both individuals and societies. This powerful impact of art is manifested through artworks. It should be noted that the concept of an artwork, an inseparable part of art, is the tangible output of the artists' creative processes, materialising as a product or performance with aesthetic values.

Benjamin (2008) argues that the uniqueness and aura of an 'artwork' are of paramount importance. In this regard, artworks must possess a unique existence and historical context. In other words, these characteristics are fundamental elements that construct the originality and historical value of the work.

The unique and historical nature of artworks, as described, transforms them from mere aesthetic objects into tools of cultural memory and social critique (Shiner, 2001). The individual and cultural elements presented by the artist during the creative process play a significant role in the interpretation and evaluation of the work. Thus, the necessity of emphasising the importance of the artist becomes apparent. Therefore, when addressing the concepts of art and artwork, it is also essential to define the concept of the artist, who is the creator of this entire process (Adajian, 2022).

'Artist' can be defined as an individual who produces and/or performs works of art. However, this general definition as 'productor and/or performer' is insufficient for defining this term. It is crucial to emphasise that the artwork expresses the emotions, thoughts, and worldview too of the individual who performs and/or creates it. Moreover, the artist not only reflects personal emotions, thoughts, and worldview but also produces or performs works that embody the impact of social issues and cultural matters too (Guoa and Guib, 2021).

The creativity and technical skills of artists nourish the originality and depth of their works. Thus, artists can powerfully reflect both their inner world and at the same time, the external world, offering a unique experience to their audience. It must be reiterated, due to its significance, that the works of artists are not merely aesthetic products but also serve as a means of communication, social critique, and cultural expression. In this context, the role of the artist extends beyond being merely a creator-performer; they can also be considered critics and cultural transmitters too (Guoa and Guib, 2021).

2.3 AI and art

The discussion above encompasses the concepts of AI and, specifically, art. In the context of the research subject, it is pertinent to delve into the relationship between AI and art.

AI can be interpreted as a force contributing to significant transformations in the art world. These transformations extend across a broad spectrum, influencing aspects ranging from the creation of artworks to their exhibition.

Starting with the use of AI in the production of artworks, learning models such as Generative Adversarial Networks (GANs) and Recursive Neural Networks (RNNs) play a crucial role in this domain (Wang et al., 2020). These learning models facilitate the integration of AI into the process of artistic creation. For instance, the work "Memories of Passersby I" by Klingemann (2018) incorporates portraits randomly selected using GANs, serving as an example that highlights the synergy between the artist's creativity and the productivity of AI.

As emphasised, the relationship between AI and art not only extends to the production of artworks but also contributes to offering new aesthetic perspectives or viewpoints to artists. Algorithms capable of analysing the structure of a particular artwork and generating new pieces serve as illustrative examples in this regard (Mazzone and Elgammal, 2019).

The types of artworks that can be produced through AI, whether visual, auditory, literary, etc., are analogous to categories of artworks that can be created by individuals (See Uzun et al., 2020).

Following the brief exploration of the relationship between AI and art, it is fitting to address the concept of copyright, which constitutes a significant focal point in the study.

2.4 Copyrights

If we approach the concept of copyright chronologically, the art creations produced by early human communities were evaluated with a different perspective than today. In ancient times, artworks were associated with materials, and they were not valued in terms of intellectual context separately from the substances they were made of. In other words, for instance, a music piece was evaluated in conjunction with the vinyl record it was pressed onto, and a painting was considered in connection with the canvas on which it was drawn. This situation eliminated the need for artists to secure themselves financially and spiritually on an individual basis (Dolunay and Keçeci, 2017).

As time progressed, the notion that artworks were essentially intellectual structures not necessarily tied to or limited by materials began to prevail among artists. At this point, the initial step is considered to be the emergence of 'printing privileges.' Especially with the rise of copying and piracy markets in the mediaeval period, the concept of 'copyright' entered human life, marking the inception of the first examples of copyright, particularly in the field of printing (Dolunay and Keçeci, 2017).

With the proliferation of copyright, the first legal regulation was the Act Anne enacted in 1709 in England (Dolunay and Keçeci, 2017).

The aforementioned concept of copyright is extremely crucial concerning the transmission of artistic production and cultural heritage from generation to generation. It is highly important for artists to find emotional and material satisfaction when presenting their works. Moreover, in a fair environment, it is vital for artists to obtain their rights in a commercial context and be able to control unauthorised usage. In this context, the protection of copyright for artworks is perceived not only as a commercial necessity but also as highly significant for cultural advancements (Özgür, 2020).

3 Traditional and digital copyrights

3.1 Traditional regulations (national: north Cyprus and international)

Geographical timeline of copyright laws in contemporary Turkey reveals a progressive journey. In the mid-1800s, the "Hukuku Telif Nizamı" (Copyright Law) was introduced, followed by the "Hakkı Telif" (Right of Copyright) enacted by the Ottoman administration in the early 1900s. The present-day legal framework, the Copyright and Related Rights Law No. 5846, was enacted on January 1, 1952 (What is copyright?, 2014).

In the current territorial boundaries of Northern Cyprus, serious challenges in copyright enforcement have led many national artists to protect their artistic works financially and spiritually by seeking copyright under the laws of the Republic of Turkey. This is due to significant issues within the copyright system in Northern Cyprus (Violation of intellectual property rights in the TRNC violation of intellectual property rights in the TRNC, 2023).

As emphasised earlier, the ever-advancing and unstoppable pace of global technology provides immense access to artists and artworks for individuals worldwide. This phenomenon has made copyright protection more challenging, leading to widespread piracy. Copyright laws play a crucial role in legally safeguarding the intellectual creations of artists. Northern Cypriot artists face considerable difficulties in protecting their artistic ideas due to the inadequacy of the existing law (Violation of intellectual property rights in the TRNC violation of intellectual property rights in the TRNC, 2023).

On a national level, the Copyright Law in Northern Cyprus is notably insufficient. The existing law, referenced from the 1911 British law during the British colonial period, is extremely limited, represented by a 4-article Law No. 264. Additionally, there are regulations such as the Broadcasting High Board Copyright and Producer Rights Protection Law and Procedures Regulation and the Copyright Regulation (Violation of intellectual property rights in the TRNC violation of intellectual property rights in the TRNC, 2023).

Under the Copyright framework applied in Northern Cyprus, the rights granted to the copyright holder include:

1 "To produce, reproduce, perform, or publish the work."
2 "In the case of a dramatic work, to adapt it into a novel or into a non-dramatic work."
3 "In the case of a novel or other non-dramatic work or other artistic works, to transform it into a dramatic work by performing it generally."
4 "In literary, dramatic, or musical works, to record, perforate with perforated waltzes, make cinematographic films, or create other devices or apparatus by which the work can be mechanically performed and represented" (Violation of intellectual property rights in the TRNC violation of intellectual property rights in the TRNC, 2023).

In the event of copyright infringement in Northern Cyprus, the copyright owner must have documentation proving ownership before initiating legal action. These cases primarily aim to prevent others from profiting from or copying and imitating creative works. It is crucial for artists to be aware that copyright protection is not indefinite, lasting for 50 years after the artist's death (Violation of intellectual property rights in the TRNC violation of intellectual property rights in the TRNC, 2023).

Examining various countries' national regulations, leading countries in copyright protection, such as Germany, the United Kingdom, and Spain, have comprehensive laws known as 'Intellectual and Artistic Works Protection Laws.' These laws protect both written and printed works as well as draft and conceptual works, highlighting the ownership of the artist (Turan, 2014).

Internationally, some crucial agreements include:

Bern Convention (1886): The oldest and most fundamental regulation governing copyright, signed in Bern, Switzerland, in 1886, providing mutual legal protection among member nations.

World Intellectual Property Organisation (WIPO) Agreement: Aims to universally regulate intellectual and artistic works' protection on a global scale, similar to the Bern Convention (Gin, 2004).

Rome Convention (1961): Primarily focuses on protecting sound recordings internationally.

TRIPS Agreement (1994): Aims to unite companies and rights holders for universal protection of intellectual property, specifically in trademark matters.

The present-day evolution of the artist and artwork, undergoing rapid change, coupled with the impact of social media, digitization, and the emergence of AI, has transformed copyright issues into a more complex and globalised landscape (Kaynak and Koç, 2015).

3.2 Traditional to digital

Traditional copyright laws have long been an effective tool for protecting the ideas and works of artists and ensuring their rights are respected. However, with the development of digital structures, social media, and art platforms, adapting copyright to digital versions has become inevitable (Litman, 2020).

Digitalised copyright, tailored to the digital environment, is considered to be more solution-oriented compared to traditional copyright methods, as it not only protects the rights of intellectual and artistic creators on an international scale but also provides fair and controllable conditions in areas such as equal access, rapid compensation processes, and work diversity (Litman, 2020).

In addition to digital platforms, the concept of art has evolved into a new dimension today, with professionally produced artworks created with AI. When examining artworks and ideas generated by AI, there is observed a divergence of opinions among individuals studying this field. Especially in the realm of art, the majority tends to accept the idea that AI is an object without personality and does not possess intellectual ownership. Therefore, artworks produced by AI are generally considered to be owned by the one directing it (Zorluel, 2019).

On the other hand, considering the analysis and production capabilities of today's computers, some argue that AI can also fulfil the emotional aspect of art creation. In this regard, attributing the intellectual ownership of artworks to AI is not surprising (Zorluel, 2019).

A group that combines and evaluates the two contrasting views mentioned above argues that the usage rights of works produced with AI are equally divided between AI and the person directing it (Zorluel, 2019; Darvishi et al., 2022).

Apart from the issue of ownership of copyright in artistic productions with AI, important aspects such as how it will be protected, etc., are not yet fully regulated. In national regulations, including the United States (Isohanni, 2021), there are no regulations in this regard. This issue, due to its current relevance, is a topic of academic discussions. It is noteworthy that even the United States, often cited as an example in these academic debates for its legal regulations, does not yet have a comprehensive legal framework on this matter (Škiljić, 2021). In particular, the international regulations mentioned above are not up-to-date in the face of today's technology. These regulations seem to be outdated, as they are based on the idea that only individuals can create works that will be protected by copyright. However, with the emergence of works produced by or with the help of AI, these regulations have become inadequate to meet current needs (Hristov, 2020).

Additionally, without straying from the focus of the topic, it is important to note that discussions regarding artworks produced through AI continue not only from a legal standpoint but also from an ethical perspective. Notably, literature includes warnings that their use in media could lead to disinformation, mass manipulation, and the proliferation of large amounts of low-quality content. However, while this is a related issue, it constitutes a separate debate. Therefore, it has been briefly addressed in this study due to its significance (Vyas, 2022).

4 Relationship between emotional state and artistic production

4.1 Emotional state

The concept of emotional state can be defined as a general evaluation of an individual's emotional condition at a given moment (Ekman and Davidson, 1994). Emotional states are categorised as positive, negative, or neutral. Emotional states reflect individuals' emotional experiences in the specified categories and manifest themselves physically, behaviourally, and mentally (Keltner and Gross, 1999).

Emotional states are closely linked to brain structures and functions. The limbic system, particularly structures such as the amygdala and hypothalamus, plays a crucial role in regulating emotional responses (LeDoux, 1996). Neurotransmitters, especially serotonin and dopamine, have a critical role in transmitting chemical signals that affect emotional states (Hariri and Holmes, 2006).

When considering the formation of emotional states, the complexity of genetic, environmental, and cognitive factors becomes apparent (Gross and Thompson, 2007). Determining factors include individuals' past experiences, general personality structure, and emotional regulation strategies acquired from childhood onwards (John and Gross, 2004).

Emotional states significantly impact individuals' quality of life. For instance, positive emotional states can enhance individuals' motivation and strengthen overall feelings of well-being (Fredrickson, 2001). However, negative emotional states can create stress and be among the factors contributing to conditions such as anxiety and even depression (Watson and Clark, 1984).

In academic studies, it is noted that individuals' emotional states are crucial; rapid and unjust progress goals can lead to unethical approaches, and the unethical use of AI is cited as an example of these unethical approaches (Dolunay and Temel, 2024).

On the other hand, considering the significance of emotional states in all areas of life, it can be argued that they are particularly important for artists when thinking about how emotional states are expressed. In this regard, it is pertinent to discuss the effects of emotional states on artistic production.

4.2 Effects on artistic production

As mentioned above, emotional states significantly impact individuals, and in this context, they also have important effects on the creativity of artists. For example, according to Csikszentmihalyi's Flow theory, positive emotional states are believed to enhance creativity and allow for deeper connections in artistic production (Moneta and Csikszentmihalyi, 1996).

Moreover, neurological studies also explore the connection between emotional states and artistic production processes. Interactions with artworks are suggested to increase activity in the emotional regions of individuals' brains (Kawabata and Zeki, 2004). Research on how emotional states affect brain activities during artistic production processes contributes to the biological understanding of the subject.

Art is considered to be a method for artists to express their emotional states through the creation of artworks (Garrido and Schubert, 2015). These processes also contribute to establishing emotional connections between viewers and art (Juslin and Sloboda, 2010).

Studies suggest that emotional states influence various processes in artists' works, from composition to colour selection (Smith et al., 2015). Similarly, studies using brain imaging techniques have found that positive emotional states support activities in brain regions associated with creativity (Flaherty, 2011).

In numerous scientific studies, the effects of emotional states on artists have been investigated, yielding varied results. As mentioned above, some studies have found that positive emotional states enhance creativity among artists (Baas et al., 2008). However, another study suggests that negative emotions also have the potential to enhance creativity (Forgas and Baumeister, 2019). Furthermore, some studies have found no significant relationship between emotional states and creativity (Davis, 2009).

While the notion that positive emotions tend to support creativity is more common, it is also plausible to assume that certain negative emotions may also support creativity and productivity. This is because some challenges, among other factors, can serve as a form of reverse psychology, providing motivation. For example, sad situations may lead the creativity (Ashby and Isen, 1999). However, the assumption that emotional states are completely unrelated to creativity is, in our view, not universally valid. Nevertheless, these differing results underscore the importance of further research and investigation into the subject.

When considering the overall consensus of consistent views, it is concluded that positive emotional states support the creativity and productivity of artists, while negative emotional states may have the opposite effect.

In this context, concerning the crucial issue of copyright for artists, if their rights are violated or they perceive that their rights have been violated, it can be assumed that their emotional states will be negatively affected. This, in turn, could undermine their creativity and motivation to create.

To emphasise the multifaceted and contemporary nature of the issue, it is noteworthy that various aspects, such as the emotional perspectives of artists towards their self-produced musical works compared to those produced with AI assistance, are also subjects of debate. Research highlighted in the literature (See Vikström and Von-Bonsdorff, 2022) provides current and intriguing insights in this direction. These studies suggest that the emotional states of artists might be negatively impacted by the notion that individuals they perceive as less talented could produce similar works with the aid of artificial intelligence (Anantrasirichai and Bull, 2022; Vikström and Von-Bonsdorff, 2022). Given that the topic is discussed from many angles, this study specifically explores the emotional states of artists regarding the issue of copyright ownership in AI-generated artworks, addressing an aspect that, in our opinion, could have significant implications for creativity and productivity.

5 Research

5.1 Method

This study, particularly focusing on the negative impacts of copyright issues on the emotional states of artists, aims to identify and propose solutions to the problems that may arise or have arisen. In this context, considering both national (Northern Cyprus) and international regulations, the emergence of copyright ownership issues in the context of the development of AI in the current digital age becomes crucial regarding the emotional states of artists. As mentioned earlier, negative emotional states can undermine creativity and artistic production, making the investigation of this issue even more significant. In the research section of this study, focus group interviews were conducted with artists and semi-structured in-depth interviews were conducted with academic experts in the field.

A focus group interview is defined as a discussion and interview conducted between a small group and a leader, utilising group dynamics to gather detailed and comprehensive information and generate ideas (Bowling, 2002; Çokluk et al., 2011). On the other hand, the primary concern in a focus group interview is to conduct a carefully planned discussion in an environment where participants can comfortably express their views and opinions (Krueger and Casey, 2000). The aim at this point is to obtain in-depth and comprehensive qualitative information on individuals' perspectives, interests, experiences, tendencies, thoughts, perceptions, feelings, attitudes, and habits regarding a specified topic. Focus group interviews typically consist of 8–12 participants (Kitzinger, 1994, 1995; Bowling, 2002; Çokluk et al., 2011).

Interviews are commonly used as a professional technique or auxiliary tool in various research fields within social sciences, including journalism, law, and medicine (Kahn, 1983; Tekin, 2006: 101). Used frequently as a data collection technique in qualitative research, interviews provide an opportunity for the participants to express themselves directly, allowing the researcher to conduct comprehensive observations about the interviewees (McCracken, 1998: 9; Tekin, 2006: 102). The interview, by asking questions covering all dimensions of the research topic, enables the collection of detailed answers, facilitating one-on-one information gathering (Johnson, 2002: 106; Tekin, 2006: 102).

Interviews can be categorised into three subtypes: unstructured, semi-structured, and structured (Punch, 2005: 166; Tekin, 2006: 104). Semi-structured interviews involve pre-prepared questions but allow the researcher to direct additional questions based on the course of the interview.

In this study, focus group interviews and semi-structured in-depth interview techniques were chosen for its relevance to the topic.

5.2 Sampling

In the context of research, the term "population" refers to the entirety of individuals with similar characteristics, while smaller groups that can be selected from within the population and have the power to represent it are referred to as "samples" (Yıldırım and Şimşek, 2018). The snowball sampling technique was employed to identify both the focus group (artists) and the groups interviewees (academic experts).

The snowball sampling technique is a method that involves selecting a reference person related to the subject of the study and reaching other individuals through recommendations. This method is iterative, and participants guide researchers, contributing to the growth of the sample. Therefore, it is known as the "snowball effect" (Biernacki and Waldorf, 1981).

The interview group identified through snowball sampling consists of eight artists (two painters, two graphic artists, two musicians, two photographers) and eight psychology experts, also two communication experts and two legal experts too.

Artists were included in the sample due to their focus on the subject. On the other hand, academic experts in communication and law were included in the sample because copyright is an interdisciplinary subject from both legal and communicative perspectives (Mengüşoğlu, 2015). On the other hand, the psychology experts were included in the sample for understanding the emotional states term, relationship between copyrights and artists emotions, emotional effects of copyrights issues on artists creativity.

While there was no geographical limitation in the context of the universality of the subject in snowball sampling, certain criteria were sought for proposing names to be included in the sample. Among these criteria were the involvement of artists in one of the fields of painting, graphics, music, or photography to create a diverse artistic perspective in the sample, having at least 5 years of professional experience to benefit from the experiences of experienced individuals, and for artists to have academic studies in the relevant fields (being academics). Similarly, for communication and legal experts, the criteria included having at least 5 years of professional experience to benefit from the experiences of experienced individuals and holding a doctoral degree in the fields of communication and law to bring an academic perspective to the subject.

On the other hand, as mentioned above, since creating a comfortable environment and careful planning are essential in focus group interviews and interviews in general, to ensure participants can openly express their opinions, the interviews were conducted online to minimise the effects of external stimuli. This approach aimed to enable focus group participants to express their views freely in their natural environments. However, considering the possibility that participants, even in their natural settings and experienced in their fields, might provide biased responses due to an overly individual perspective, academic artist-scholars with professional academic experience and titles were included in the focus group. This inclusion aimed to obtain more objective, ethical, and unbiased responses from an academic perspective.

Similarly, the primary purpose of selecting field experts (psychologists, legal experts, and communication specialists) from among academics with professional experience and titles at universities was to maintain objectivity and academic rigour. As focus group interviews typically consist of 8–12 participants, it was deemed sufficient to conduct a focus group interview with eight artist-scholars relevant to the subject. Since the focus group comprised eight individuals and given the equal importance of the psychologists in the context of the study—as professional evaluators—interviews were conducted with eight academic psychologists. While the groups of artists and psychologists each consisted of eight members, the groups of legal scholars and communication scholars consisted of only two members each. The reasoning behind this was to avoid deviating from the main focus of the study while still supporting it from legal and communicative perspectives, as copyright issues encompass both areas.

Details regarding the analysis of the data obtained from focus group interviews and semi-structured in-depth interviews are provided below.

5.3 Analyses

The data obtained in the research were evaluated using the content analysis method. Content analysis is the neutral, systematic, and quantitative description of the content resulting from communication (Berelson, 1952: 17; Koçak and Arun, 2006: 22). Another definition characterises content analysis as a research technique used to derive repeatable and valid results from data (Krippendorff, 1986: 25).

According to yet another definition, content analysis is a research technique where valid interpretations extracted from the text are articulated through successive processes (Weber, 1989: 5; Koçak and Arun, 2006: 22).

In line with the nature of the study, sometimes detailed coding is required, while at other times, comprehensive coding may not be necessary (Karataş, 2017: 80; Yıldırım and Şimşek, 2018: 233). Therefore, due to the ease of evaluating the data obtained through the conducted interviews in the context of the study, a more intricate coding and categorisation theme was not deemed necessary.

In this context, the data obtained from focus group and semi-structured in-depth interviews with sampling were themed and coded as follow:

In the analyses under the specified themes and codes (Table 1), 20% of the direct opinions of the interview group were included. The names of the participants are given in codes as A1, A2, A3, etc. (for Artists), P1, P2, P3, etc. (for Psychology Experts), L1, L2 (for Law Experts) and C1, C2 (for Communication Experts).

5.4 The importance and general status of copyright

5.4.1 Importance of copyright

The matter of copyright is considered a fundamental value in European countries and holds significant importance within the legal domain. Particularly, due to the rapid development of digital platforms in recent times, copyright has been subject to serious infringements, necessitating concrete and ultimate legal remedies through robust legislation and judiciary, as emphasised by Eren (2019). Within this regards, it is believed that copyright laws lacking strong and solid foundations lead to serious problems and infringements within the context of artists.

The interview group (artists, communication experts and law experts), entirely in line with the literature, highlighted the importance of copyright and expressed that artists' rights over their works are protected within the framework of copyright:

> A6: *"Many globally recognized artists attach great importance to the copyright of their artworks, both in terms of financial and spiritual gains."*
>
> L2: *"From a legal perspective, copyright laws are the most effective regulations that protect artists internationally in terms of profit and ownership on global platforms."*

5.4.2 General status

Copyright laws constitute a legal framework related to the rights granted to the creators of creative works. Broadly defined, copyright entails the producer's exclusive rights over a work of art, including but not limited to copying, distribution, profit generation, and transfer to others. While national copyright laws may vary, international copyright agreements uphold fundamental principles.

TABLE 1 Themes and codes.

	Themes				
	Copyright: importance, general status	Emotional states, and creativity	Regulations of copyright (national and international)	The impact of copyright infringements on artists' emotional states and copyright ownership in AI-generated content production	Inadequacy of regulations in the context of copyright issues and proposed solutions
Codes	Importance of Copyright	Emotional States	National (North Cyprus) Regulations	The Relationship between Copyright Issues and Emotional States	Inadequacy of Regulations of Copyrights
	General Status	Emotional States and Creativity	International Regulations	Copyright Ownership in AI-Generated Content Production	Proposed Solutions
				The Emotional Impact on Artists When AI Holds Exclusive Copyright	

In alignment with this perspective, the interview group (artists, communication experts and law experts) also asserted that copyright, in the context of its general status, relies on a legal infrastructure:

> C1: *"...laws and legal authorities should be in place to prevent the infringement of copyright."*
>
> A2: *"...taking Turkey, our closest geographical neighbour, as an example, copyright is protected under the law, just as it is worldwide, through the Intellectual and Artistic Works Act."*

5.5 Emotional states, and creativity

5.5.1 Emotional states

Emotional states significantly impact individuals' quality of life. For instance, positive emotional states can enhance individuals' motivation and strengthen overall feelings of well-being (Fredrickson, 2001). However, negative emotional states can create stress and be among the factors contributing to conditions such as anxiety and even depression (Watson and Clark, 1984).

In the context of expertise, during the interview, only psychologists were asked about what emotions are. Psychologists characterised emotions as fundamental elements that can fluctuate and directly impact individuals.

> P1: *"Emotions are processes that can vary and influence an individual's mood positively or negatively."*
>
> P3: *"A state of feeling that an individual experiences over an extended period."*

5.5.2 Emotional states and creativity

In literature, some studies have found that positive emotional states enhance creativity among artists (Baas et al., 2008). Another perspective suggests that in some cases, emotional states like sadness may trigger a reverse psychology effect, influencing creativity (Ashby and Isen, 1999).

> *P2: "Emotional states are directly linked to creativity. Positive emotional states enhance creativity and productivity, whereas negative emotional states can hinder them. However, in some cases, for instance, an artist in a melancholic state may still produce a creative work. Yet, generally, positive emotional states will more strongly support creativity and productivity."*
>
> *P4: "Emotional state influences creativity and productivity. Generally, positive emotional states will have a positive impact on generating and creativity."*

5.6 Regulations of copyright (national and international)

5.6.1 National (north Cyprus) regulations

In the context of national copyright matters, the Copyright Law in Northern Cyprus is notably inadequate, mirroring the deficiency observed in many laws within the region. Furthermore, it is worth noting that Northern Cyprus has not independently formulated the law currently in effect within its borders. Influenced by the island's historical hosting of various civilizations and the dominance of the United Kingdom before Turkish sovereignty, the existing Copyright Law, Chapter 264, consisting of only four articles, refers to the 1911 law of the UK, yet remains highly insufficient (Violation of intellectual property rights in the TRNC violation of intellectual property rights in the TRNC, 2023).

The study group (artists, communication experts and law experts), unanimously, expressed that the existing Copyright Regulations at the national level are inadequate:

> A3: *"I am aware that the rights of use for works of art are not sufficiently protected in Cyprus. We have encountered several difficulties, especially in terms of commercial aspects of artistic works."*
>
> L1: *"Copyrights are not protected in our country."*

5.6.2 International regulations

Universal copyright regulations that are deemed significant include:

Berne Convention (1886): Historically, this is the most rooted and ancient regulation within the framework of copyright. Signed in the city of Bern, Switzerland, in 1886, the convention establishes fundamental approaches and provides mutual legal protection among member nations. It aims to safeguard the rights of artistic works concerning ideas and ownership (Berne Convention, 1886).

World Intellectual Property Organisation (WIPO) Agreement: The WIPO agreement is a global regulation that addresses works in terms of both intellectual and ownership aspects. Similar to the Berne Convention mentioned above, it aims to protect international intellectual and artistic works.

Rome Convention (1961): Primarily focusing on sound recordings, the Rome Convention defends the rights of printed sound works on an international scale.

TRIPS Agreement (1994): The TRIPS (Trade-Related Aspects of Intellectual Property Rights) Agreement aims to unite companies and intellectual property holders in protecting copyright, particularly in the context of branding, for those who are members.

Within the interview group, 50% of the respondents referred to the aforementioned agreements in response to the relevant question:

> L1: *"Copyrights are internationally safeguarded by agreements. For example, the Rome Convention, the Berne Convention are among the important regulations...."*

5.7 The impact of copyright infringements on artists' emotional states and copyright ownership in AI-generated content production

5.7.1 The relationship between copyright issues and emotional states

Art is considered a method of expressing the emotional states of artists, the process of producing art, and the emotions of the artists themselves (Garrido and Schubert, 2015). These processes also contribute to establishing emotional connections between viewers and art (Juslin and Sloboda, 2010).

Studies suggest that emotional states influence various aspects of artists' works, from composition to colour selection (Smith et al., 2015). Furthermore, research utilising brain imaging techniques has identified that positive emotional states support activities in brain regions associated with creativity (Flaherty, 2011).

In the context of emotional states, psychologists with their expertise on copyright issues have stated that the emotions of artists whose copyrights are violated will be affected.

> P5: *"The emotional states of artists whose copyrights are violated will be adversely affected."*
>
> P6: *"Artists whose rights or directly copyrights are violated will experience adverse effects both financially and emotionally, impacting their motivation to create and consequently, their creativity."*

5.7.2 Copyright ownership in AI-generated content production

In the literature, there is an ongoing debate regarding the ownership of copyright for artworks generated by AI. According to the first perspective, copyright ownership should belong to the AI, while the second perspective argues that it should be attributed to the individual who directs and contributes intellectual effort to the AI. The third perspective suggests that copyright should be jointly owned by both the AI and the individual.

The interview group (artists, communication experts and law experts), approximately ≅83%, expressed the view that copyright should be equally shared between AI and the individual:

> A5: *"Since a work of art is created through the collaborative efforts of both the artist and the AI program, the copyright ownership should be divided equally."*
>
> A8: *"In a collaborative work where both parties contribute, the distribution of copyright should be equal."*

On the other hand, approximately ≅8% of the interview group stated that only the individual should hold the copyright, while another ≅8% argued for the sole copyright ownership of the AI.

5.7.3 The emotional impact on artists when AI holds exclusive copyright

As mentioned earlier, the emotional states of artists significantly influence their productivity and creativity (Flaherty, 2011). If artists cannot claim copyright for a work they have put effort into producing, it is possible that they may be emotionally affected negatively.

The questions at this point were directed towards artists and psychology experts in their respective fields to understand emotional states. Approximately ≅87.5% of the interview group expressed that artists would not be satisfied if copyright for works produced using AI belonged solely to the AI. They emphasised that such a situation could lead to negative emotional and creative consequences for artists:

> A1: *"I believe that the inability of artists to claim rights for artworks produced from their own ideas will lead to reluctance in the production process."*
>
> P7: *"The absence of copyright on a work that an artist has put effort into will negatively impact their emotional state, reducing their desire to create."*

On the other hand, around ≅12.5% expressed the opinion that the emotional states of artists should not be affected if only AI holds the copyright.

5.8 Inadequacy of regulations in the context of copyright issues and proposed solutions

5.8.1 Inadequacy of regulations of copyrights

In North Cyprus, national copyright laws and regulations are known to be inadequate. Even if copyright is obtained for works within the framework of these laws, the protection of this copyright will not be universally recognised on an international platform since the country is not acknowledged (See Tamçelik, 2013; Dolunay and Kasap, 2020). In this regard, making an international agreement and raising awareness among artwork owners through state-supported programmes will play a crucial role in preventing unauthorised developments (Dolunay and Keçeci, 2017).

The interview group, based on the conducted interviews, unanimously expressed that the existing copyright law in effect in Northern Cyprus is insufficient:

> A7: *"In my opinion, our artists' artworks are not adequately protected. In fact, in many cases, artists find themselves having to personally safeguard their works."*
>
> L2: *"Due to legal gaps, our artists are in a very disadvantaged situation on a national level. They often fall victim to intellectual property theft, and unfortunately, there is no well-organised protective law they can rely on to assert their rights."*

5.8.2 Proposed solutions

Intellectual property regulations have been established based on the notion that only individuals can create works eligible for copyright. However, as indicated, these regulations have become inadequate in addressing current needs, particularly concerning works generated by or through AI and the associated copyright issues (Hristov, 2020).

The interview group expressed similar views, highlighting the insufficiency of international regulations in meeting contemporary needs, especially in the copyright processes of artworks created with AI. Consequently, they proposed solutions in this regard.

The interview group provided various recommendations to prevent or address copyright issues. Each participant responded with multiple suggestions, broadly categorised under two main headings:

In the context of the importance of copyright, proposing widespread educational activities to increase respect for and awareness of copyright:

> A2: *"Certain institutions need to educate individuals on this matter. I believe that individuals involved in the arts, seeking copyright for their artworks, should be obligated to attend an initial training."*

Development of national and international legal regulations related to copyright / Establishment of regulations for digital copyright and works generated through AI:

> A7: *"I believe that traditional copyright laws have reached their limits, both in national and universal contexts. I am of the opinion that all copyright issues need to be resolved in the context of digital technology, ensuring easy accessibility and a lasting solution."*

6 Findings and discussion

Copyright, a crucial system enabling the protection of artists' rights over their works, is grounded in its legal status. As with any right, it is essential to respect and adhere to legal regulations concerning this right.

In the study, the interview group emphasised the importance of copyright and indicated that its status is based on a legal foundation when defining copyright. This approach aligns with the perspective in the literature, emphasising the significance of copyright within the framework of copyright law (Vaver, 2000).

Emotional state, often relying on long-term accumulation, can typically exhibit moment-to-moment fluctuations. It is a concept used to express individuals' positive, negative, or neutral feelings (Fredrickson, 2001). In this study, it was determined that emotional states could indeed be positive, negative, or neutral, consistent with this notion in the literature.

In the literature, highlights an association between emotional states and creativity, suggesting that positive emotional states generally positively influence creativity and the desire to produce (See Baas et al., 2008). Consistent with this approach, this study found that emotional states are linked to productivity and creativity, and the positive emotional states increase creativity and productivity.

On the other hand, the common belief that negative emotional states adversely affect creativity and productivity was considered in this study and presented as a finding too. However, there is also an approach in the literature suggesting that some negative emotional states (such as sadness), through a reverse psychology effect, may support creativity and productivity (Ashby and Isen, 1999).

It is emphasised once again that while positive emotional states positively influence creativity and productivity, negative emotional states generally have an adverse effect on aspects of creativity and productivity for artists. However, with the perspective mentioned above, it was also found in this study's research findings that certain negative emotional states, such as sadness, might positively trigger creativity and productivity through a reverse psychology effect (as exceptions).

The regulation of copyright, stemming from its legal foundation, is made possible through various legal arrangements. In the context of this study, the issue is regulated nationally (in Northern Cyprus) by the Copyright Law Chapter 264. On the international front, crucial agreements such as the Berne Convention (1886), the World Intellectual Property Organisation (WIPO) Agreement, the Rome Convention (1961), and the TRIPS Agreement (1994) regulate the issue.

In the study, the interview group also mentioned the laws and agreements mentioned above concerning national (Northern Cyprus) and international regulations.

Furthermore, considering the importance and legal regulations, the connection between copyright and the emotional states of artists is crucial. The concept of emotional states can be defined as an individual's overall assessment of their emotional condition at a given moment (Ekman and Davidson, 1994). Emotional states can be categorised as positive, negative, or neutral. Positive emotional states can enhance individuals' motivation and overall feelings of well-being (Fredrickson, 2001). On the other hand, negative emotional states can create stress and contribute to conditions such as anxiety and even depression (Watson and Clark, 1984).

In the study, it was found that if artists encounter copyright issues, as expected, there will be negative effects on their emotional states. Negative emotional states can not only affect artists' overall well-being but also have a negative impact on their desire to create and their creativity.

When examining the process in the context of AI, it was observed that the issue of copyright ownership arises concerning works produced through AI. In the literature, there are three different approaches to the copyright of artworks produced through AI: 1. ownership to AI, 2. ownership to the individual or individuals directing and contributing to the work, 3. joint ownership shared by both (Darvishi et al., 2022).

All these views lead to separate discussions. The suggestion that AI does not have a separate personality and only the individual should have copyright ownership, the proposal for joint ownership, or the idea that AI, as the creator, should have copyright are suggestions that need to be examined in detail.

In the study, the interview group mostly stated that the individual's intellectual effort and guidance are crucial. However, if the production is carried out with the help of AI, they emphasised that copyright should be in partnership between the individual and AI.

On the other hand, regarding the views mentioned above, the study found that the copyright belonging to AI would negatively affect the emotional states of artists. This situation could negatively impact artists' desires to create and their creativity.

In a national context (Northern Cyprus), the existing copyright law consisting of four articles was found to be outdated, not only regarding digital copyright but even traditional copyright. It leaves artists without basic protection.

On an international scale, although there are international agreements regarding traditional copyright, it was stated that new agreements are needed, particularly addressing digital copyright and the copyright of works produced through AI.

This assertion is consistent with the claim in the literature that national and international regulations are insufficient, especially in terms of the copyright of artworks produced through AI (Hristov, 2020).

In the study, the interview group proposed solutions for preventing or resolving copyright issues:

- Emphasising the importance of awareness-raising educational activities nationally and internationally.
- Suggesting that the national copyright law (in Northern Cyprus) needs to be revised to protect traditional copyright and be expanded to include digital copyright, especially for works produced through AI.

– On an international level, emphasising the need to revise international agreements to include regulations for works produced through AI or to create a new agreement based on the importance of this issue.

7 Instead of conclusion: copyrights, AI, emotions and future?

All rights are significant; however, considering the nourishing, developmental, and transformative potential of culture and art on individuals, society, and the world, the importance of copyright can be emphasised.

Copyright, grounded in a legal framework, is a fundamental system that safeguards artists' rights over their works. While traditional copyright is regulated by national laws and international agreements, the contemporary digital age has introduced AI as a significant player.

Before addressing issues related to AI, it is crucial to reiterate that the violation of copyright can adversely affect the emotional states of artists, subsequently impairing their creative desires and abilities.

On the other hand, the debate over copyright ownership in works produced through AI—whether it should belong to the artist, the AI, or be shared jointly—requires careful consideration as a current and sensitive issue.

A fair approach may involve preserving ownership for the artist in the presence of their intellectual effort and unique direction, while also granting some level of rights to the AI as the creative tool. This situation may lead to various problems and debates, highlighting the need for a fair legal framework. Furthermore, legal and just considerations should be accompanied by an awareness of the emotional states that artists may experience. For instance, a scenario where only AI holds the copyright, while potentially fair legally, may negatively impact the emotional well-being of artists.

In a national context, specifically in Northern Cyprus, even traditional copyright has not been adequately regulated. In this regard, traditional copyright should be prioritised for regulation, and any legal framework should encompass artworks produced through AI, considering the demands of the contemporary era.

On the international stage, there is a clear need for the expeditious implementation of legal regulations concerning artworks produced through AI, aligning with the demands of the times. The continuously evolving technology has rendered existing international regulations inadequate.

Emphasising the importance of copyright nationally and internationally, educational activities aimed at raising awareness among artists and the general public should be organised by universities at the national level and international intellectual property organisations.

For future academic endeavours, psychologists are recommended to conduct in-depth interviews, including aspects of artists' emotional states affected by the violation of their rights, to delve deeper into the subject. Furthermore, as highlighted in this paper, which reexamines and analyses the effects of evolving technology, it is crucial to identify and discuss the benefits and drawbacks of the digital age. In this context, it is recommended that academic research systematically investigating the use of AI in artistic works be conducted in collaboration between artists and AI experts. Additionally, given the interdisciplinary nature of the topic, it is suggested that communication, ethics, and legal experts collaborate to deeply explore the communicative, ethical, and regulatory aspects of the issue. Communication specialists should address the communicative dimensions, ethics experts should examine the ethical considerations, and legal professionals should investigate the regulatory aspects, ultimately developing concrete proposals for communicative, ethical, and legal frameworks.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by the Near East University Scientific Research Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

HK: Formal analysis, Methodology, Writing – original draft, Writing – review & editing. AD: Formal analysis, Methodology, Writing – original draft, Writing – review & editing, Conceptualization, Project administration, Supervision.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adajian, T. (2022) in The Definition of Art, The Stanford Encyclopedia of Philosophy. ed. E. N. Zalta. Available at: https://plato.stanford.edu/archives/spr2022/entries/art-definition/

Adorno, W. T. (1997). Aesthetic Theory. London: The Athlone Press Ltd.

Akdoğan, B. (2001). Sanat, Sanatçı, Sanat Eseri ve Ahlak [art, artist, artwork and morality]. *J. Faculty of Divinity of Ankara University* 42, 1–245. doi: 10.1501/ILHFAK_0000000533

Akipek, Ş., and Dardağan, E. (2001). Law applicable to copyright violations occurred in the virtual environment. *Ankara University Faculty of Law J.* 50, 1–139. doi: 10.1501/Hukfak_0000000589

Anantrasirichai, N., and Bull, D. (2022). Artificial intelligence in the creative industries: a review. *Artif. Intell. Rev.* 55, 589–656. doi: 10.1007/s10462-021-10039-7

Ashby, F. G., and Isen, A. M. (1999). A neuropsychological theory of positive affect and its influence on cognition. *Psychol. Rev.* 106, 529–550. doi: 10.1037/0033-295X.106.3.529

Baas, M., De Dreu, C. K., and Nijstad, B. A. (2008). A meta-analysis of 25 years of mood-creativity research: hedonic tone, activation, or regulatory focus? *Psychol. Bull.* 134, 779–806. doi: 10.1037/a0012815

Balkır, S. (2020). Art-artist and art work as a Meta object. *J. Arts* 3, 31–44. doi: 10.31566/arts.3.004

Barasch, M. (2013). Theories of art: 1. From Plato to Winckelmann. New York: Routledge.

Benjamin, W. (2008). The Work of Art in the Age of Its Technological Reproducibility, and Other Writings on Media. London: Harvard University Press.

Bengio, Y. (2021). The malicious use of artificial intelligence: forecasting. *Prevention Mitigation.* doi: 10.48550/arXiv.1802.07228

Berelson, B. (1952). Content analysis in communication research. Glencoe: Free Press.

Biernacki, P., and Waldorf, D. (1981). Snowball sampling: problems and techniques of chain referral sampling. *Sociol. Methods Res.* 10, 141–163. doi: 10.1177/004912418101000205

Bowling, A. (2002). Research methods in health: Investigating health and health services. Philadelphia, PA: McGraw-Hill House.

Çokluk, Ö., Yılmaz, K., and Oğuz, E. (2011). A qualitative interview method: focus group interview. *J. Theoretical Educ. Sci.* 4, 95–107.

Darvishi, K., Liu, L., and Lim, S. (2022). Navigating the Nexus: legal and economic implications of emerging technologies. *Law Econ.* 16, 172–186. doi: 10.35335/laweco.v16i3.59

Davis, M. A. (2009). Understanding the relationship between mood and creativity: a Meta-analysis. *Organ. Behav. Hum. Decis. Process.* 108, 25–38. doi: 10.1016/j.obhdp.2008.04.001

Dolunay, A. (2024). The use of artificial intelligence in the field of communication: a research on the perspectives of communication academics. *J. Autonomous Intelligence* 15, 1–10. doi: 10.32629/jai.v7i5.1610

Dolunay, A., and Kasap, F. (2020). Still unrecognized state "Turkish republic of northern Cyprus" in the context of the Cyprus negotiations: status of the TRNC' court decisions. *J. Politics Law* 13, 1–9. doi: 10.5539/jpl.v13n3p1

Dolunay, A., and Keçeci, G. (2017). Copyright problems in the Turkish Cypriot law within the framework of communication ethics. *J. History Culture Art Res.* 6, 1396–1409. doi: 10.7596/taksad.v6i4.1081

Dolunay, A., and Temel, A. C. (2024). The relationship between personal and professional goals and emotional state in academia: a study on unethical use of artificial intelligence. *Front. Psychol.* 15:1363174. doi: 10.3389/fpsyg.2024.1363174

Dreyfus, H. L. (1972). What computers Can't do: The limits of artificial intelligence. USA: Harper & Row.

Ekman, P., and Davidson, R. J. (1994). The nature of emotion: Fundamental questions. London: Oxford University Press.

Eren, C. S. (2019). Striking a balance between freedom of expression and copyrights-case-law of the European courts. *İnsan Hakları Yıllığı* 37, 18–41.

Erinç, S. M. (1988). Art and education. *J. Uludağ University Faculty of Educ.* 3, 175–182.

Flaherty, A. W. (2011). Brain illness and creativity: mechanisms and treatment risks. *Can. J. Psychiatr.* 56, 132–143. doi: 10.1177/070674371105600303

Forgas, J. P., and Baumeister, R. F. (2019) in The social psychology of gullibility conspiracy theories, fake news and irrational beliefs. eds. M. A. Runco and S. R. Pritzker (New York: Routledge).

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: the broaden-and-build theory of positive emotions. *Am. Psychol.* 56, 218–226. doi: 10.1037/0003-066X.56.3.218

Garrido, S., and Schubert, E. (2015). Moody melodies: do they cheer us up? A study of the effect of sad music on mood. *Psychol. Music* 43, 244–261. doi: 10.1177/0305735613501938

Gillotte, J. (2020). Copyright infringement in AI-generated artworks. *UC Davis Law Rev.* 53, 2655–2691.

Gin, E. (2004). International copyright law: beyond the WIPO & TRIPS debate. *J. Patent and Trademark Office Society* 86:763.

Gross, J. J., and Thompson, R. A. (2007). "Emotion regulation: conceptual foundations" in *Handbook of emotion regulation*. ed. J. J. Gross, vol. *298* (New York: The Guilford Press), 1805–1824.

Guoa, Z., and Guib, J. (2021). Definition of Artists. *International Journal of Frontiers in Sociology* 3, 8–10. doi: 10.25236/IJFS.2021.030902

Haiven, M. (2015). Art and money: three aesthetic strategies in an age of financialisation. *Finance and Society* 1, 38–60. doi: 10.2218/finsoc.v1i1.1370

Hariri, A. R., and Holmes, A. (2006). Genetics of emotional regulation: the role of the serotonin transporter in neural function. *Trends Cogn. Sci.* 10, 182–191. doi: 10.1016/j.tics.2006.02.011

Hristov, K. (2020). Artificial intelligence and the copyright survey. *J. Sci. Policy Gover.* 16, 1–18.

Isohanni, P. (2021). Copyright and human originality in artistic works made using artificial intelligence, Aalto University School of Business, Master Thesis.

John, O. P., and Gross, J. J. (2004). Healthy and unhealthy emotion regulation: personality processes, individual differences, and life span development. *J. Pers.* 72, 1301–1334. doi: 10.1111/j.1467-6494.2004.00298.x

Johnson, J. (2002). In-Depth Interviewing. Handbook of Interview Research Context & Method (Editors: Jaber F. Gubrium, James A. Holstein). London: Sage Publications.

Juslin, P. N., and Sloboda, J. A. (2010). "The past, present, and future of music and emotion research" in *Handbook of music and emotion: Theory, research, applications* (Oxford University Press), 933–955.

Kahn, R. L. (1983). The dynamics of interviewing. Florida: Robert E. Krieger Publishing Company.

Karataş, Z. (2017). Paradigm transformation in social sciences research: Rise of qualitative approach. *Turkish Journal of Social Work Research* 1, 68–86.

Kawabata, H., and Zeki, S. (2004). Neural correlates of beauty. *J. Neurophysiol.* 91, 1699–1705. doi: 10.1152/jn.00696.2003

Kaynak, S., and Koç, S. (2015). New challenges in copyright law: social media. *Folklore/Literary* 21, 389–410.

Keltner, D., and Gross, J. J. (1999). Functional accounts of emotions. *Cognit. Emot.* 13, 467–480. doi: 10.1080/026999399379140

Kitzinger, J. (1994). The methodology of focus groups: the importance of interaction between research participants. *Sociology of Health and Illness* 16, 103–121. doi: 10.1111/1467-9566.ep11347023

Kitzinger, J. (1995). Qualitative research: introducing focus groups. *Br. Med. J.* 311, 299–302. doi: 10.1136/bmj.311.7000.299

Klingemann, M. (2018). Memories of Passersby I. Retrieved from: https://underdestruction.com/2018/12/29/memories-of-passersby-i

Koçak, A., and Arun, Ö. (2006). The Sampling Problem in Content Analysis Studies. *Journal of Selcuk Communication* 3, 21–28.

Krippendorff, K. (1986). Content Analysis: An Introduction to its Methodology: Sage.

Krueger, R. A., and Casey, M. A. (2000). Focus groups: A practical guide for applied research. California: Sage.

Kurzweil, R. (2005). The singularity is near: When humans transcend biology. USA: Viking Press.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

LeDoux, J. E. (1996). The emotional brain: The mysterious underpinnings of emotional life. USA: Simon & Schuster.

Litman, J. (2020). Real copyright reform. New York: Prometheus Books.

Mazzone, M., and Elgammal, A. (2019). Art, creativity, and the potential of artificial intelligence. *Art* 8, 1–9. doi: 10.3390/arts8010026

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Mag.* 27:12. doi: 10.1609/aimag.v27i4.1904

McCracken, G. (1998). The long interview. California: Sage Publications.

McCulloch, W. S., and Pitts, W. (1943). A logical Calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics.* 5, 115–133. doi: 10.1007/BF02478259

Mengüşoğlu, T. (2015). Felsefi Antropoloji Bakımından Tecrübe Mefhumunun Tahlili (analysis of the concept of experience in philosophical anthropology). *Felsefe Arkivi* 3, 136–161.

Moneta, G. B., and Csikszentmihalyi, M. (1996). The effect of perceived challenges and skills on the quality of subjective experience. *J. Pers.* 64, 275–310. doi: 10.1111/j.1467-6494.1996.tb00512.x

Newell, A., and Simon, H. (1956). The logic theory machine: a complex information processing system. *IRE Transactions on Information Theory* 2, 61–79. doi: 10.1109/TIT.1956.1056797

Nilsson, N. J. (2010). Artificial intelligence: A new synthesis, vol. *17*. California: Morgan Kaufmann, 57–63.

Özgür, A. (2020). Art and copyright. *J. Culture Art Res.* 9, 101–114.

Punch, K. F. (2005). Introduction to social research: quantitative and qualitative approaches. eds. D. Bayrak, H. B. Aslan and Z. Akyüz Ankara: Siyasal Publishing.

Russell, S., and Norvig, P. (2022). Artificial intelligence: A modern approach. *4th* Edn. Pearson England: Pearson Series.

Shiner, E. L. (2001). The Invention of Art: A Cultural History. Chicago: The Univesity of Chicago Press.

Škiljić, A. (2021). When art meets technology or vice versa: key challenges at the crossroads of AI-generated artworks and copyright law. *Int. Rev. Intellectual Property and Competition Law* 52, 1338–1369. doi: 10.1007/s40319-021-01119-w

Smith, S., Nichols, T., and Vidaurre, D. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nat. Neurosci.* 18, 1565–1567. doi: 10.1038/nn.4125

Tamçelik, S. (2013). The properties of some resolutions adopted by the un security council regarding Cyprus and their analytical evaluation (1964-1992). *Turk. Stud.* 8, 1229–1268. doi: 10.7827/TurkishStudies.6041

Tekin, H. H. (2006). In-depth interview of qualitative research method as a data collection technique. *Istanbul University J. Sociol.* 3, 101–116.

Turan, M. (2014). Types of works on law on intellectual and artistic works: a comparative analysis. *Information World* 15, 125–158.

Turgut, İ. (1991). Sanat Felsefesi [philosophy of art]. İzmir: Bilgehan Publishing.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind* LIX, 433–460. doi: 10.1093/mind/LIX.236.433

Uzun, Y., Akkuzu, B., and Kayrıcı, M. (2020). The relationship of artificial intelligence to culture and art. *European J. Sci. Technol.* 28, 753–757. doi: 10.31590/ejosat.1010691

Vaver, D. (2000). Copyright law. Toronto: Irwin Law.

Vikström, C., and Von-Bonsdorff, M. (2022). Changes in musicians' perceptions and feelings as their original compositions are altered by AI. (dissertation). Retrieved from https://kth.diva-portal.org/smash/record.jsf?pid=diva2%3A1704343&dswid=2304

Violation of intellectual property rights in the TRNC violation of intellectual property rights in the TRNC. (2023). Retrieved from: https://l24.im/XpFAW

Vyas, B. (2022). Ethical implications of generative AI in art and the media. *Int. J. Multidis. Res.* 4, 1–11. doi: 10.36948/ijfmr.2022.v04i04.9392

Wang, L., Chen, W., Yang, W., Bi, F., and Yu, F. R. (2020). A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE* 8, 63514–63537. doi: 10.1109/ACCESS.2020.2982224

Watson, D., and Clark, L. A. (1984). Negative affectivity: the disposition to experience aversive emotional states. *Psychol. Bull.* 96, 465–490. doi: 10.1037/0033-2909.96.3.465

Weber, R. P. (1989). Basic content analysis. Londra: Sage.

What is copyright? (2014). Retrieved from: http://www.telifhaklari.gov.tr/Telif-Hakki-Nedir

Yıldırım, A., and Şimşek, H. (2018). Qualitative research methods in the social sciences. *11th* Edn. Ankara: Seçkin Publishing.

Zorluel, M. (2019). Artificial intelligence and the copyright. *Union of Turkish Bar Assoc. Rev.* 142, 305–356.

frontiers | Frontiers in Psychology

TYPE Systematic Review
PUBLISHED 12 August 2024
DOI 10.3389/fpsyg.2024.1391832

OPEN ACCESS

EDITED BY
Simone Belli,
Complutense University of Madrid, Spain

REVIEWED BY
Isabella Poggi,
Roma Tre University, Italy
Alan D. A. Mattiassi,
University of Modena and Reggio Emilia, Italy

*CORRESPONDENCE
Carla Susinna
✉ carla.susinna@irccsme.it

RECEIVED 26 February 2024
ACCEPTED 18 July 2024
PUBLISHED 12 August 2024

CITATION
Morgante E, Susinna C, Culicetto L, Quartarone A and Lo Buono V (2024) Is it possible for people to develop a sense of empathy toward humanoid robots and establish meaningful relationships with them? *Front. Psychol.* 15:1391832.
doi: 10.3389/fpsyg.2024.1391832

Is it possible for people to develop a sense of empathy toward humanoid robots and establish meaningful relationships with them?

Elena Morgante, Carla Susinna*, Laura Culicetto, Angelo Quartarone and Viviana Lo Buono

IRCCS Centro Neurolesi Bonino Pulejo, Messina, Italy

Introduction: Empathy can be described as the ability to adopt another person's perspective and comprehend, feel, share, and respond to their emotional experiences. Empathy plays an important role in these relationships and is constructed in human–robot interaction (HRI). This systematic review focuses on studies investigating human empathy toward robots. We intend to define empathy as the cognitive capacity of humans to perceive robots as equipped with emotional and psychological states.

Methods: We conducted a systematic search of peer-reviewed articles using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines. We searched Scopus, PubMed, Web of Science, and Embase databases. All articles were reviewed based on the titles, abstracts, and full texts by two investigators (EM and CS) who independently performed data collection. The researchers read the full-text articles deemed suitable for the study, and in cases of disagreement regarding the inclusion and exclusion criteria, the final decision was made by a third researcher (VLB).

Results: The electronic search identified 484 articles. After reading the full texts of the selected publications and applying the predefined inclusion criteria, we selected 11 articles that met our inclusion criteria. Robots that could identify and respond appropriately to the emotional states of humans seemed to evoke empathy. In addition, empathy tended to grow more when the robots exhibited anthropomorphic traits.

Discussion: Humanoid robots can be programmed to understand and react to human emotions and simulate empathetic responses; however, they are not endowed with the same innate capacity for empathy as humans.

KEYWORDS

empathy, human–robot interaction, humanoid robots, social robots, rehabilitation

1 Introduction

Empathy is a multidimensional construct used to describe the sharing of another person's feelings and the ability to identify with others and grasp their subjective experiences (Airenti, 2015). It covers a spectrum of phenomena, ranging from experiencing feelings of concern for others to feeling within oneself the feelings of others. This ability is a complex phenomenon that includes an affective component understood as the capacity to share the emotional status of other subjects, and a cognitive dimension that implies the ability to rationally understand the thoughts, feelings, and perspectives of others (Decety and Jackson, 2004; Eisenberg and Eggum, 2009; Decety and Ickes, 2011).

In other words, emotional empathy enables individuals to be influenced by the emotions of others, aiding in the recognition of one's own and the interlocutor's emotions, which allows them to create a mental representation of the thoughts and emotional states of their interlocutors (Leite et al., 2013). Empathy is an extremely adaptable and versatile process that permits social behavior in a variety of settings. Although it can be considered a specific feature of humans, prosocial actions brought about by empathy can occasionally be constrained by external circumstances. Hoffman (2001) showed that constraints on empathy stem from two primary factors: empathic over-arousal and interpersonal dynamics between the subject and the target of empathy. Empathic over-arousal materializes if indications of distress are exceptionally strong; in this case, the empathic concern shifts to a state of personal distress. Moreover, the nature of the relationship between the observer and the object of empathy significantly shapes the form of the prosocial actions undertaken by the observer. For instance, people are more likely to empathize with friends and relatives than strangers (Krebs, 1970). Empathic responses can be modulated by personal characteristics or situational contexts (De Vignemont and Singer, 2006).

At the neural level, studies on empathy-mediated processes have demonstrated the important role of networks involved in action simulation and mentalizing, depending on the information available in the environment. This neural network of empathy includes the anterior insula, somatosensory cortex, periaqueductal gray, and anterior cingulate cortex (Engen and Singer, 2013).

In recent years, neuroscientific approaches have increased the study of different forms of empathy in human–robot interaction (HRI) (Tapus et al., 2007; Riek et al., 2010). This field is expanding rapidly as robots become increasingly adept at sophisticated social skills (Vollmer et al., 2018). Humanoid robots have sociable abilities and the capacity to interact with humans to understand verbal and non-verbal communication, such as postures and gestures (Alves-Oliveira et al., 2019).

Humanoid robots can influence users' emotional states and perceptions of social interactions (Saunderson and Nejat, 2019). Studies have explored how people attribute intentions, personality, and emotional meaning to robots, thus helping establish guidelines for designing more humane and engaging robotic interfaces. Using neuroimaging techniques, it is demonstrated that observation of human movements and observation of robotic movements activate the same brain areas, indicating that the anthropomorphic qualities of robots can elicit empathic responses in humans (Gazzola et al., 2007). This emphasizes the role of the mirror neuron system in regulating human empathy and imaginative processes. Mirror neurons facilitate not only the reproduction of observed actions but also emotional resonance with others. This system responds not only to human actions but also activates in response to actions performed by a robot (Iacoboni, 2009).

The robots understand human intentions using the properties of the mirror neuron system, and they may be able to more accurately anticipate human actions and respond to it more precisely (Han and Kim, 2010).

Empathy is viewed as an active body of ongoing emotional and cognitive exchanges rather than a singular phenomenon of emotional mirroring to develop a relationship between individuals and other agents over time.

Research on virtual humans and robots referred to as "advanced intelligent systems" when combined explores one of two main perspectives: (1) how humans empathize with advanced intelligent systems or (2) the impact of a robot's empathetic behavior on humans.

The first viewpoint looks at how humans emotionally engage with robots that have human-like characteristics, and it does not necessarily need robots to be empathic. As for the second viewpoint, many academics have looked at different methods and algorithms to give robots empathy so they can recognize and respond to humans' emotional states (Birmingham et al., 2022).

This bidirectional empathy can strengthen the bonds between humans and robots and improve the quality of interaction and trust.

This systematic review is focused on studies that investigated empathy in the HRI.

2 Materials and methods

2.1 Search strategy

We conducted a systematic review to investigate the construct of empathy in HRI. A literature review was performed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines by searching PubMed, Web of Science, and Embase. We considered articles published between 2004 and 2023. The following key terms were used: ('empathy'[MeSH Terms] OR 'empathy'[All Fields]) AND ('humans'[All Fields] OR 'humans'[MeSH Terms] OR 'humans'[All Fields] OR 'human'[All Fields]) AND ('robot'[All Fields] OR 'robot s'[All Fields] OR 'robotically'[All Fields] OR 'robotics'[MeSH Terms] OR 'robotics'[All Fields] OR 'robotic'[All Fields] OR 'robotization'[All Fields] OR 'robotized'[All Fields] OR 'robots'[All Fields]) AND (fha[Filter]). Only English texts were considered.

All articles were reviewed based on titles, abstracts, and full texts by two investigators (EM and CS) who independently performed data collection to reduce the risk of bias (i.e., bias of missing results, publication bias, time lag bias, and language bias). The researchers read the full-text articles deemed suitable for the study, and in cases of disagreement regarding the inclusion and exclusion criteria, the final decision was made by a third researcher (VLB). The list of articles was then refined for relevance, revised, and summarized, with the key topics identified from the summary based on the inclusion and exclusion criteria.

The inclusion criteria were as follows: (i) studies on the population of healthy adults and (ii) studies that included a psychometric assessment of empathy.

The exclusion criteria were as follows: (i) studies involving children and (ii) case reports and reviews.

2.2 Data extraction and analysis

Following the full-text selection, data extraction from the included studies was summarized in a table (Microsoft Excel—Version 2021). The summarized data included the assigned ID number, study title, year of publication, first author, study aims and design, study duration, method and setting of recruitment, inclusion and exclusion criteria, informed consent, conflicts of interest and funding, type of intervention and control, number of participants, baseline characteristics, type of outcome, time points for assessment, results, and key conclusions.

The agreement between the two reviewers (CS and EM) was assessed using the kappa statistic. The kappa score, with an accepted threshold for substantial agreement set at >0.61, reflected excellent concordance between the reviewers. This criterion ensured a robust evaluation of inter-rater reliability, emphasizing the achievement of a substantial level of agreement in the data extraction process.

3 Results

A total of 484 articles were identified, including 89 from PubMed, 335 from Web of Science, and 60 from Embase. All articles were evaluated based on title, abstract, full text, and topic specificity. Only 11 studies met the inclusion criteria (Table 1; Figure 1).

Recognizing or anticipating how people will react to robots and how well robots will respond to humans may depend on an understanding of human empathy toward them.

Two distinct research areas address the topic of empathy in social robots. In the first, the human interlocutor is the observer of the robot, and the robot is the target of human empathy. In the second area investigated, the robot is the observer of the human and is designed to exhibit empathy to the human. Thus, in the selected studies, we found two empathy-based HRI design orientations: the expression of empathy and the induction of empathy. The expression of empathy means that humans feel that the social robot is empathizing with their emotions. Relative to empathy induction, the social robot expresses its emotions in advance, through which the interacting human feels empathy. In recent years, significant progress has been made in both areas.

3.1 Empathic encoded robot response

In a study by Birmingham et al. (2022), a robot that used affective empathic statements was perceived as having more empathy in comparison with a robot programmed to manifest cognitive empathy. To evaluate the interaction, participants completed a short survey after watching two demonstration videos of each condition. The study analyzed the relationship between the participants' attitudes toward the robots, their assessment of how genuine they felt the interaction was, and their assessment of the robot's empathy in each condition. Furthermore, the relevant finding concerns the participants' belief that the interaction between the robot and the actors was credible, natural, and genuine. A few studies have focused on the human characteristics of human robots, such as facial expressions, which play an important role in social interactions and communication processes. In detail, Mollahosseini et al. (2018) studied the benefits of using an automatic facial expression recognition system in the spoken dialog of a social robot and how the robot's sympathy and empathy would be affected by the accuracy of the system. In the experimental condition, the robot empathizes with the user through a series of predefined conversations. The results of the study indicate that the incorporation of an automatic facial expression recognition system allowed subjects to perceive the robot as more empathetic than in the other conditions. In a study by Leite et al. (2013), two players engaged in a chess game were accompanied by an autonomous robot expressing empathy. In this way, the robot acted as a social companion. In this study, the empathic behaviors reported in the literature were modeled in a social robot capable of inferring certain affective states of the human subject, reacting emotionally to these states, and commenting appropriately on a chess game. The results indicate that individuals toward whom the robot behaved empathetically perceived the robot as friendlier, which continues to support the hypothesis that empathy plays a key role in HRI. These findings serve to support investigations concerning HRI focusing on human emotions and the development of robots that are perceived as appropriately empathic and that can tailor their empathic responses to users.

3.2 Robot-dependent empathic human response

Moon et al. (2021) studied empathy induction, which outlined the appropriate emotional expressions for a social robot to elicit empathy-based behavior. Like human–human interactions, non-verbal cues have been found to significantly influence empathy and induced behavior when people interact with robots. Specifically, the results showed that non-verbal cues conveyed a negative emotion, appropriate to the situation; this had a decisive effect on perceived emotion, empathy, and behavior induction. It has also been shown that a robot's affective narrative can also influence its ability to elicit empathy in human subjects. In the study by Frederiksen et al. (2022), the authors explore the stimulation of empathy by investigating interaction scenarios involving a robot that uses affective narratives to generate compassion in subjects, while failing to complete the task. Therefore, this study explores the relationship between the type of narrative conveyed by the robot (funny, sad, and neutral) and the robot's ability to elicit empathy in interactions with human observers. The results demonstrate that the type of narrative approach of the robot was able to influence the level of empathy created during an interaction. Konijn and Hoorn (2020) compared the facial articulacy of humanoid robots to a human in affecting users' emotional responsiveness, showing that detailed facial articulacy does make a difference. The results of the study showed that robots can arouse empathic reactions in humans; when these reactions are greater, the robot's facial expression will be more complex. The expressiveness of the robot has an important communicative function and makes it usable in contexts such as healthcare and education, allowing users to affectively relate to the robot at a level appropriate to the task or objective. Corretjer et al. (2020) focused on the study of quantitative indicators of early empathy realization in a challenging scenario, highlighting how participation in a collaborative activity (solving a maze) between humans and robots influenced the development of empathy. In a subsequent study (Corretjer et al., 2020), they assessed empathy, using indicators such as affective attachment, trust, and expectation regulation. Through the development of these aspects in an atmosphere that is supportive, the

TABLE 1 Studies assessing empathy.

Study	Aim	Sample (*N*)	Empathy assessment	Robotic agent	Outcomes
Empathic robot					
Birmingham et al. (2022)	Examining how viewers perceive cognitive and affective empathetic statements from a robot in response to human disclosure	111 Healthy Subjects	RoPE scale, modified from a first-person questionnaire to a third-person questionnaire	Empathic Agent	The participants rated the affective statements higher than the cognitive ones
Mollahosseini et al. (2018)	Assessing automated FER accuracy on robots interacting with humans, along with task engagement, empathy, and likability	16 Healthy Subjects	Designed Questionnaire	Ryan Companionbot	Participants rated the empathic robot higher in empathy and likability compared to non-empathic robot
Leite et al. (2013)	Assessing whether empathetic artificial companions enhance user relationships	40 Healthy Subjects	Designed Questionnaire	iCat	Participants rated the supportive robot higher in companionship, alliance, and self-validation
Empathic responses of humans					
Tsumura and Yamada (2022)	The text explores whether human empathy varies based on task difficulty and content	578 Healthy Subjects	12-item questionnaire modified from the IRI	Empathic Agent	Higher task difficulty promoted human affective empathy
Konijn and Hoorn (2020)	The significance of facial articulacy and emotions in optimizing human–robot communication	265 Healthy Subjects	Designed Questionnaire	Robot Alice and robot Nao/ Zora	Humans showed less empathic and emotional responsiveness toward robots compared to humans
García-Corretjer et al. (2023)	Active collaboration enhances meaningful empathy between humans and robots	18 Healthy Subjects	Toronto Empathy Questionnaire (TEQ)	Robot Robobo	Participants trusted the robot's suggestions amid uncertainty, demonstrating teamwork attitudes
Erel et al. (2022)	Examining whether non-humanoid robot gestures boost emotional support in human–human interaction	64 Healthy Subjects	Verbal Empathy	Non-humanoid robotic object	Robots performing empathetic gestures improve human emotional support interaction
Spaccatini et al. (2023)	Assessing how attributing a specific mind to a social robot affects empathy toward individuals in distress	269 Healthy Subjects	Online questionnaire on Qualtrics	Social Robot's Anthropomorphism, Chatbot	The level of anthropomorphization of robots produces empathy in interaction with humans
Moon et al. (2021)	Studying non-verbal cues' impact and mediation structure in human–robot interaction	48 Healthy Subjects	Designed Questionnaire	Social Robot 'Hubot'	A non-verbal cue has an outweighing effect on empathy in HRI
Frederiksen et al. (2022)	Studying how a robot's emotional story affects empathy in humans	220 Healthy Subjects	Designed Questionnaire	Kuri robot	Sad narrative increased participants' empathy and willingness to help the robot
Corretjer et al. (2020)	To develop empathy through fun collaboration scenario in which a user and a social robot work together	10 Healthy Subjects	Designed Questionnaire	Robot Robobo	Developing empathy through engaging collaboration scenarios with a social robot

IRI, interpersonal reactivity index; RoPE, robot's perceived empathy; FER, facial expression recognition.

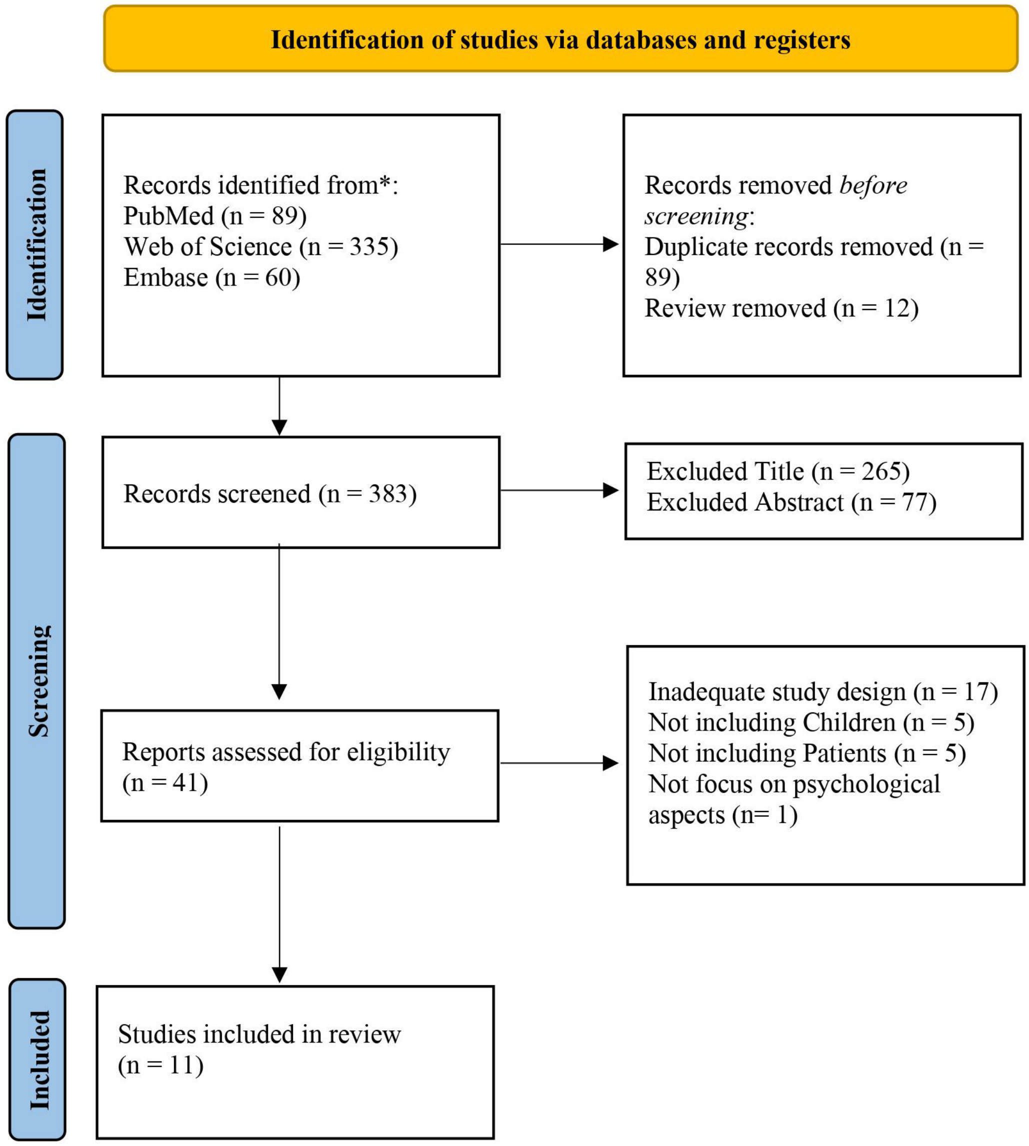

Identification of studies via databases and registers
Identification
Records identified from*:
PubMed (n = 89)
Web of Science (n = 335)
Embase (n = 60)
Records removed *before screening*:
Duplicate records removed (n = 89)
Review removed (n = 12)
Screening
Records screened (n = 383)
Excluded Title (n = 265)
Excluded Abstract (n = 77)
Reports assessed for eligibility (n = 41)
Inadequate study design (n = 17)
Not including Children (n = 5)
Not including Patients (n = 5)
Not focus on psychological aspects (n= 1)
Included
Studies included in review (n = 11)

FIGURE 1
Search and selection of eligible articles.

participants in the study engaged in mutual understanding, listening, reflecting, and performing. Although the robots did not have anthropomorphic characteristics, the participants managed to establish a collaborative and empathetic relationship with them with the aim of achieving a common goal. Tsumura and Yamada (2022), in an experimental condition, studied the conditions required to develop empathy toward anthropomorphic agents. The findings demonstrated that greater task difficulty, independent of task content, increased human empathy toward robots. Spaccatini et al. (2023) examined the potential impact of anthropomorphized robots on human social perceptions. The authors induced anthropomorphization of social robots by manipulating the level of anthropomorphism of their appearance and behavior. The results demonstrated that anthropomorphic social robots were associated with higher levels of experience and agency. Furthermore, the type of mind attributed to the anthropomorphic social robot influences the empathy perceived by the human. Erel et al. (2022) have shown that the non-verbal gestures of a non-humanoid robot can increase emotional support in human–human interactions. This indicates that a robot even without anthropomorphic features can improve the way humans interact.

4 Discussion

Many studies on people's empathy for robots have been published in the last few years, but there are also fundamental questions concerning the correct use of the term empathy (Niculescu et al., 2013; Darling et al., 2015; Seo et al., 2015). Generally, empathy can be described as the ability to comprehend and experience another person's feelings and experiences and is a crucial component of human social interaction that promotes the growth of affection and social bonds (Anderson and Keltner, 2002). When considering humanoid robots, one may wonder whether people can develop empathy for a device (Malinowska, 2021).

The phenomenon of humans' empathy toward robots has garnered significant attention in the field of HRI and is in some ways a controversial topic. As reported in numerous studies, empathy in HRI is bidirectional. On the one hand, humans can feel empathy toward robots; on the other, robots, with the progress of technology, are designed to be empathetic in interactions with humans. It is possible to feel empathy toward robots, especially when the latter possess human characteristics, are anthropomorphized (Breazeal, 2003; Paiva et al., 2017), and adopt human-like attitudes. When robots exhibit human-like facial expressions, gestures, or voices, people tend to perceive them as more relatable and emotionally expressive, which can trigger empathetic reactions (Riek et al., 2010).

Social robots with human-like features can affect how people feel about them, which in turn can impact the robots' ability to convey emotions (Spaccatini et al., 2023). The modulation of voice tone has also been shown to be effective in promoting empathic processes (James et al., 2018). In addition, robots designed with expressive faces that can mimic sadness, happiness, or surprise are more likely to elicit empathetic responses from humans (Leite et al., 2013). Humans tend to see robots with human-like characteristics as more than just machines, attributing them with a sense of liveliness and even emotional capabilities. This can lead people to perceive anthropomorphic robots as companions, promoting acceptance and trust between humans (Zoghbi et al., 2009). Consequently, people are more likely to interpret the emotions expressed by such robots as genuine, which can facilitate emotional connections in human–robot interaction (Bartneck et al., 2007).

It was also examined how mirroring facial expressions could improve empathy in HRI. Robots capable of reproducing human facial expressions seem to significantly improve empathic engagement (Gonsior et al., 2011).

However, while giving robots human-like features can enhance their ability to express emotions and help people understand those emotions, it can also lead to a phenomenon known as the "Uncanny Valley (Mori, 2005; Misselhorn, 2009)." This effect describes a decrease in human empathy toward robots and an increase in discomfort as robots become more similar to humans (Mori et al., 2012). Based on several studies, it has been discovered that this effect occurs in environments with a high level of anthropomorphism and various sensory stimuli, including auditory, visual, and tactile cues (Nomura and Kanda, 2016). As a result, individuals may develop incorrect expectations of the robot's cognitive and social abilities during prolonged interactions (Dautenhahn and Werry, 2004).

Several studies have shown that humans' empathic involvement toward robots can extend to various situations, even those in which robots are perceived as being in difficulty or in need of help. In such a scenario, it has been seen that people may feel guilt or sadness when they observe a robot failing to complete a task or being mistreated (Darling et al., 2015).

The anthropomorphism of robots might influence the socio-cognitive processes of humans and the subsequent behavior of subjects toward them. In the study by Spatola and Wudarczyk (2021), a focus was placed on the emotional capabilities of the robot, pointing out that endowing robots with more complex emotions could lead to more anthropomorphic attributions toward them. Therefore, the perceived emotionality of robots, which is not limited to one type of emotion, could predict some of the characteristics of robot anthropomorphism (Schömbs et al., 2023).

This assumption is in line with the "Simulation Theory" which suggests that the way we understand the minds of others is through "simulating" the situation of another; therefore, it should be more immediate to empathize with the emotions and mental states of a robotic agent that has human characteristics (Mattiassi et al., 2021).

Even non-humanoid robots are capable of activating empathic responses; in fact, they can produce behaviors and responses that users perceive as social or emotional, promoting the development of empathy. For example, if a robot has been programmed to provide help or comfort, users are more likely to feel empathy toward it, regardless of its non-human physical characteristics (Erel et al., 2022).

As robotics and artificial intelligence continue to advance, integrating empathic capabilities into robots has emerged as a crucial area of research. Empathy, the ability to understand and respond to the emotions of others, is fundamental to human social interaction. Developing an empathetic robot, like any other robot, requires a clear definition of its purpose. Based on this purpose, designers can create interaction scenarios, and engineers can develop the robot's software and hardware architecture (Park and Whang, 2022). Transferring this capability to robots promises to revolutionize various fields, including healthcare, education, and social care, by improving the quality and effectiveness of human–robot engagement (Johanson et al., 2023). For social agents to exhibit empathic behavior autonomously, they need to simulate the empathic processes; indeed, empathic robots are designed to recognize, interpret, and respond appropriately to human emotions, thus promoting more natural and meaningful interactions. These robots have the potential to provide companionship, support therapeutic interventions, and assist in the care of vulnerable populations, such as the elderly or people with special needs (Darling et al., 2015).

However, humanoid robots have significant limitations in terms of empathy. Humanoid robots cannot participate in social relationships, as they are defined in the empathic mode, because they do not satisfy the requirements of logical and purposeful subjectivity. A being with logical subjectivity can think, reason, and make decisions independently based on his or her own understanding. This concept means that, despite technological advances and progress in the empathic design of robots, they have limitations: Robots do not yet possess autonomous cognitive processes and therefore lack logic and intentionality. Their responses, while potentially sophisticated and human-like, are ultimately the result of programmed behaviors rather than authentic understanding or shared emotional experiences.

While they can recognize emotions such as sadness or anger, they have difficulty understanding the underlying causes or motivations. This is the prerequisite for true empathy, which requires not only recognizing emotions but also sharing and understanding the feelings involved. Most robots cannot feel real emotions on their own. They can simulate emotional reactions, but these are based on algorithms and data and not on real feelings (Chuah and Yu, 2021). Researchers on HRI have begun to investigate various aspects of empathy in robots. Understanding and feeling the emotions of another human person requires a high level of emotional awareness and understanding that current systems do not possess. Humanoid robots can be made to understand and respond to human emotions using pre-programmed algorithms and models. They can be programmed to simulate empathetic responses to some degree extent for certain applications in HRI and social robotics, but they do not have the same innate capacity for empathy as human people (Johanson et al., 2021). However, numerous studies in the field of HRI have shown that humans may empathize with and trust robots that can recognize their emotional states and respond appropriately to them (Kozima et al., 2004).

Regarding mental state perception/attribution, which is the cognitive ability to reflect on one's own and others' mental states such as beliefs, desires, feelings, and intentions by robots, studies have described contrasting results. While, on the one hand, people attribute the behavior of robots to underlying mental causes, on the other, they tend to deny that robots have a mind when explicitly requested to do so (Thellman et al., 2022).

While, on the one hand, people attribute the behavior of robots to underlying mental causes, on the other, they tend to deny that robots have a mind when explicitly requested to do so.

The bias of people to attribute mental states to robots is the outcome of multiple factors, including the motivation, behavior, appearance, and identity of robots. Endowing them with mental states helps to predict and explain their behavior, reduces uncertainty, and increases the sense of control in an interaction context (Epley et al., 2007; Eyssel et al., 2011; Levin et al., 2013; de Graaf and Malle, 2019). Indeed, it has also been found that people are more likely to attribute mental states to robots both when they are designed to exhibit socially interactive behavior and when they are endowed with a human-like appearance.

In most studies in the literature, it appears that the theory of mental state attribution is most often related to anthropomorphism, i.e., the attribution of mental abilities and human traits to non-human entities (Thellman et al., 2022).

Social robots are an increasingly important component of an improved social reality with relationships. Although true empathy in humanoid robots may still be a long way off, recent advances in social and developmental psychology, neuroscience, and virtual agent research have shown promising avenues for the development of empathic social robots (Guzzi et al., 2018). Seibt (2017) has classified different levels and degrees of sociality in human–robot interactions within the social interactions framework (SISI) and used the concept of 'simulation' to distinguish between full realization, partial realization, and different simulated forms of social processes, such as approximation, representation, imitation, mimicry, or replication. SISI can simulate some aspects of this complexity, but it cannot fully replicate the real-time dynamics and emotional subtleties of real human interactions (Seibt, 2017).

The main limitation of this review is the significant weakness in defining empathy, as it is not a directly observable construct but can only be inferred from behavior, and there is no clear definition or global agreement on how to measure empathic abilities in robots.

Human beings' attributions of robots are related to dimensions of mental perception. These depend on both experience and behavior and suggest that the more mental state attribution capabilities are ascribed to robots, the more they are likely to be valued (Gray et al., 2007).

Furthermore, the overall quality of evidence was low and the selected studies differed greatly in their definitions, assessment tools, and outcome measures. Due to the lack of standardized protocols, a meta-analysis could not be conducted. Regarding the assessment of perceived empathy, the way humans empathize with robots can be measured by their behavior toward robots (Spatola and Wudarczyk, 2021). Empathic emotions can be expressed through facial expressions, bodily expressions, physiological reactions, and action tendencies, and then through explicit measures such as surveys (Carpinella et al., 2017), and currently also through neuroscientific measures (e.g., EEG, MRI, and fNIRS). Although various questionnaires are available to study empathy in humans, in particular Davis' questionnaire (Davis, 1983), which is undoubtedly a benchmark for measuring individual differences in empathy, many researchers have developed their measures without relying exclusively on the currently existing instruments. The main controversy in assessment concerns the fact that to assess robot-induced empathy, one must rely on human subjects' perception of empathic traits, which means that one must measure the degree of 'perceived empathy'. The evaluation has a major impact on future developments and on whether more emphasis should be placed on certain algorithms or certain functional constructs rather than others. Therefore, evaluations also provide data that will influence the creation of new models for robot behavior, which in turn will affect the many different new applications.

The implications of empathy in HRI are manifold. Another important aspect is understanding why humans feel empathy toward robots as this influences the design and effectiveness of these interactions (Leite et al., 2014; Stock-Homburg, 2022). The goal of researchers must be to develop new design models to increase the emotional intelligence and social integration of robots and ultimately create more effective and realistic human–robot interactions (Damiano et al., 2015).

Improving these interactions must both increase the quality of the user experience and have beneficial therapeutic outcomes. Despite promising applications, the development of truly empathetic robots is fraught with complex challenges, including ethical implications. While empathy enhances human–robot interactions, it also raises ethical questions about the nature of these interactions and the potential for emotional manipulation (Coeckelbergh, 2010). To improve the utility and acceptance of robots in society, future perspectives must also consider these implications and ensure that robots are designed to promote positive and healthy human–robot interactions without exploiting human emotions (Zhou and Shi, 2011).

Data availability statement

The data presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

EM: Methodology, Writing – original draft. CS: Methodology, Writing – original draft. LC: Writing – review & editing. AQ: Supervision, Writing – review & editing. VL: Conceptualization, Methodology, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by Current Research Fund 2024 Ministry of Health, Italy.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Airenti, G. (2015). The cognitive bases of anthropomorphism: from relatedness to empathy. *Int. J. Soc. Robot.* 7, 117–127. doi: 10.1007/s12369-014-0263-x

Alves-Oliveira, P., Sequeira, P., Melo, F. S., Castellano, G., and Paiva, A. (2019). Empathic robot for group learning. *ACM Trans. Hum. Robot. Interact* 8, 1–34. doi: 10.1145/3300188

Anderson, C., and Keltner, D. (2002). The role of empathy in the formation and maintenance of social bonds. *Behav. Brain Sci.* 25, 21–22. doi: 10.1017/S0140525X02230010

Bartneck, C., Kanda, T., Ishiguro, H., and Hagita, N. (2007). Is the Uncanny Valley an uncanny cliff?. RO-MAN 2007- The 16th IEEE International Symposium on Robot and Human Interactive Communication, Jeju, Korea (South), pp. 368–373.

Birmingham, C., Perez, A., and Matarić, M. (2022). Perceptions of cognitive and affective empathetic statements by socially assistive robots. In 2022 17th ACM/IEEE international conference on human-robot interaction (HRI). 323–331. IEEE.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *Int. J. Hum. Comput. Stud.* 59, 119–155. doi: 10.1016/S1071-5819(03)00018-1

Carpinella, C. M., Wyman, A. B., Perez, M. A., and Stroessner, S. J. (2017). The robotic social attributes scale (Rosas) development and validation. In proceedings of the 2017 ACM/IEEE international conference on human-robot interaction, pp. 254–262.

Chuah, S. H. W., and Yu, J. (2021). The future of service: the power of emotion in human-robot interaction. *J. Retail. Consum. Serv.* 61:102551. doi: 10.1016/j.jretconser.2021.102551

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics Inf. Technol.* 12, 209–221. doi: 10.1007/s10676-010-9235-5

Corretjer, M. G., Ros, R., Martin, F., and Miralles, D. (2020). The maze of realizing empathy with social robots. In 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). 1334–1339.

Damiano, L., Dumouchel, P., and Lehmann, H. (2015). Towards human–robot affective co-evolution overcoming oppositions in constructing emotions and empathy. *Int. J. Soc. Robot.* 7, 7–18. doi: 10.1007/s12369-014-0258-7

Darling, K., Nandy, P., and Breazeal, C. (2015). Empathic concern and the effect of stories in human-robot interaction. In 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). 770–775.

Dautenhahn, K., and Werry, I. (2004). Towards interactive robots in autism therapy: background, motivation and challenges. *Pragmat. Cogn.* 12, 1–35. doi: 10.1075/pc.12.1.03dau

Davis, M. H. (1983). Measuring individual differences in empathy: evidence for a multidimensional approach. *J. Pers. Soc. Psychol.* 44, 113–126. doi: 10.1037/0022-3514.44.1.113

De Graaf, M. M., and Malle, B. F. (2019) People's explanations of robot behavior subtly reveal mental state inferences. In 2019 14th ACM/IEEE international conference on human-robot interaction (HRI) (pp. 239–248). IEEE.

De Vignemont, F., and Singer, T. (2006). The empathic brain: how, when and why? *Trends Cogn. Sci.* 10, 435–441. doi: 10.1016/j.tics.2006.08.008

Decety, J., and Ickes, W. (Eds.) (2011). The social neuroscience of empathy: Mit press.

Decety, J., and Jackson, P. L. (2004). The functional architecture of human empathy. *Behav. Cogn. Neurosci. Rev.* 3, 71–100. doi: 10.1177/1534582304267187

Eisenberg, N., and Eggum, N. D. (2009). Empathic responding: sympathy and personal distress. *Social Neurosci. Emp.* 6, 71–830. doi: 10.7551/mitpress/9780262012973.003.0007

Engen, H. G., and Singer, T. (2013). Empathy circuits. *Curr. Opin. Neurobiol.* 23, 275–282. doi: 10.1016/j.conb.2012.11.003

Epley, N., Waytz, A., and Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychol. Rev.* 114, 864–886. doi: 10.1037/0033-295X.114.4.864

Erel, H., Trayman, D., Levy, C., Manor, A., Mikulincer, M., and Zuckerman, O. (2022). Enhancing emotional support: the effect of a robotic object on human–human support quality. *Int. J. Soc. Robot.* 14, 257–276. doi: 10.1007/s12369-021-00779-5

Eyssel, F., Kuchenbrandt, D., and Bobinger, S. (2011). Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In Proceedings of the 6th international conference on Human-robot interaction (pp. 61–68).

Frederiksen, M. R., Fischer, K., and Matarić, M. (2022). Robot vulnerability and the elicitation of user empathy. In 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). 52–58.

García-Corretjer, M., Ros, R., Mallol, R., and Miralles, D. (2023). Empathy as an engaging strategy in social robotics: a pilot study. *User Model. User-Adap. Inter.* 33, 221–259. doi: 10.1007/s11257-022-09322-1

Gazzola, V., Rizzolatti, G., Wicker, B., and Keysers, C. (2007). The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *NeuroImage* 35, 1674–1684. doi: 10.1016/j.neuroimage.2007.02.003

Gonsior, B., Sosnowski, S., Mayer, C., Blume, J., Radig, B., Wollherr, D., et al. (2011). Improving aspects of empathy and subjective performance for HRI through mirroring facial expressions: IEEE, RO-MAN, Atlanta, GA, USA. 350–356.

Gray, H. M., Gray, K., and Wegner, D. M. (2007). Dimensions of mind perception. *Science* 315:619

Guzzi, J., Giusti, A., Gambardella, L. M., and Di Caro, G. A. (2018). A model of artificial emotions for behavior-modulation and implicit coordination in multi-robot systems. In Proceedings of the genetic and evolutionary computation conference. 21–28.

Han, J. H., and Kim, J. H. (2010) Human-robot interaction by reading human intention based on mirror-neuron system. Nel 2010 IEEE international conference on robotics and biomimetics (pp. 561–566). IEEE.

Hoffman, M. (2001). Empathy and moral development: Implications for caring and justice: Cambridge, UK, Cambridge University Press.

Iacoboni, M. (2009). Imitation, empathy, and mirror neurons. *Annu. Rev. Psychol.* 60, 653–670. doi: 10.1146/annurev.psych.60.110707.163604

James, J., Watson, C. I., and Mac Donald, B. (2018). Artificial empathy in social robots: an analysis of emotions in speech. In 27th IEEE International symposium on robot and human interactive communication (RO-MAN). 632–637.

Johanson, D. L., Ahn, H. S., and Broadbent, E. (2021). Improving interactions with healthcare robots: a review of communication behaviours in social and healthcare contexts. *Int. J. Soc. Robot.* 13, 1835–1850. doi: 10.1007/s12369-020-00719-9

Johanson, D., Ahn, H. S., Goswami, R., Saegusa, K., and Broadbent, E. (2023). The effects of healthcare robot empathy statements and head nodding on trust and satisfaction: a video study. *ACM Trans. Hum. Robot. Interact* 12, 1–21. doi: 10.1145/3549534

Konijn, E. A., and Hoorn, J. F. (2020). Differential facial articulacy in robots and humans elicit different levels of responsiveness, empathy, and projected feelings. *Robotics* 9:92. doi: 10.3390/robotics9040092

Kozima, H., Nakagawa, C., and Yano, H. (2004). Can a robot empathize with people? *Artif Life Robot.* 8, 83–88. doi: 10.1007/s10015-004-0293-9

Krebs, D. (1970). Altruism: an examination of the concept and a review of the literature. *Psychol. Bull.* 73, 258–302. doi: 10.1037/h0028987

Leite, I., Castellano, G., Pereira, A., Martinho, C., and Paiva, A. (2014). Empathic robots for long-term interaction. *Int J Soc. Robot.* 6, 329–341. doi: 10.1007/s12369-014-0227-1

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., and Paiva, A. (2013). The influence of empathy in human–robot relations. *Int. J. Hum. Comput. Stud.* 71, 250–260. doi: 10.1016/j.ijhcs.2012.09.005

Levin, D. T., Killingsworth, S. S., Saylor, M. M., Gordon, S. M., and Kawamura, K. (2013). Tests of concepts about different kinds of minds: predictions about the behavior of computers, robots, and people. *Hum. Comput. Interact.* 28, 161–191. doi: 10.1080/07370024.2012.697007

Malinowska, J. K. (2021). What does it mean to empathise with a robot? *Minds Mach.* 31, 361–376. doi: 10.1007/s11023-021-09558-7

Mattiassi, A. D., Sarrica, M., Cavallo, F., and Fortunati, L. (2021). What do humans feel with mistreated humans, animals, robots, and objects? Exploring the role of cognitive empathy. *Motiv. Emot.* 45, 543–555. doi: 10.1007/s11031-021-09886-2

Misselhorn, C. (2009). Empathy with inanimate objects and the Uncanny Valley. *Minds Mach.* 19, 345–359. doi: 10.1007/s11023-009-9158-2

Mollahosseini, A., Abdollahi, H., and Mahoor, M. H. (2018). Studying effects of incorporating automated affect perception with spoken dialog in social robots. In 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (783–789).

Moon, B. J., Choi, J., and Kwak, S. S. (2021). " Pretending to be okay in a sad voice: social Robot's usage of verbal and nonverbal Cue combination and its effect on human empathy and behavior inducement. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (854–861).

Mori, M. (2005). On the uncanny valley. Proceedings of the Humanoids-2005 workshop: Views of the Uncanny Valley. Tsukuba, Japan.

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The Uncanny Valley [from the field]. *IEEE Robot. Autom. Magaz.* 19, 98–100. doi: 10.1109/MRA.2012.2192811

Niculescu, A., van Dijk, B., and Nijholt, A. (2013). Making social robots more attractive: the effects of voice pitch, humor and empathy. *Int. J. Soc. Robot.* 5, 171–191. doi: 10.1007/s12369-012-0171-x

Nomura, T., and Kanda, T. (2016). Rapport–expectation with a robot scale. *Int. J. Soc. Robot.* 8, 21–30. doi: 10.1007/s12369-015-0293-z

Paiva, A., Leite, I., Candeias, A., Martinho, C., and Prada, R. (2017). Empathy in virtual agents and robots: a survey. *ACM Transact. Interact. Intell. Syst.* 7, 1–40. doi: 10.1145/2912150

Park, S., and Whang, M. (2022). Empathy in human–robot interaction: designing for social robots. *Int. J. Environ. Res. Public Health* 19:1889. doi: 10.3390/ijerph19031889

Riek, L. D., Paul, P. C., and Robinson, P. (2010). When my robot smiles at me: enabling human-robot rapport via real-time head gesture mimicry. *J. Multim. User Interfac.* 3, 99–108. doi: 10.1007/s12193-009-0028-2

Saunderson, S., and Nejat, G. (2019). How robots influence humans: a survey of nonverbal communication in social human–robot interaction. *Int. J. Soc. Robot.* 11, 575–608. doi: 10.1007/s12369-019-00523-0

Schömbs, S., Klein, J., and Roesler, E. (2023). Feeling with a robot—the role of anthropomorphism by design and the tendency to anthropomorphize in human-robot interaction. *Front. Robot. AI* 10:1149601. doi: 10.3389/frobt.2023.1149601

Seibt, J. (2017). "Towards an ontology of simulated social interaction: varieties of the "as if" for robots and humans" in Sociality and normativity for robots. Studies in the Philosophy of Sociality. eds. R. Hakli and J. Seibt (Springer: Cham, Switzerland).

Seo, S. H., Geiskkovitch, D., Nakane, M., King, C., and Young, J. E. (2015). Poor thing! Would you feel sorry for a simulated robot? A comparison of empathy toward a physical and a simulated robot. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (125–132).

Spaccatini, F., Corlito, G., and Sacchi, S. (2023). New dyads? The effect of social robots' anthropomorphization on empathy towards human beings. *Comput. Hum. Behav.* 146:107821. doi: 10.1016/j.chb.2023.107821

Spatola, N., and Wudarczyk, O. A. (2021). Implicit attitudes towards robots predict explicit attitudes, semantic distance between robots and humans, anthropomorphism, and prosocial behavior: from attitudes to human–robot interaction. *Int. J. Soc. Robot.* 13, 1149–1159. doi: 10.1007/s12369-020-00701-5

Stock-Homburg, R. (2022). Survey of emotions in human–robot interactions: perspectives from robotic psychology on 20 years of research. *Int. J. Soc. Robot.* 14, 389–411. doi: 10.1007/s12369-021-00778-6

Tapus, A., Mataric, M. J., and Scassellati, B. (2007). Socially assistive robotics [grand challenges of robotics]. *IEEE Robot. Autom. Magaz.* 14, 35–42. doi: 10.1109/MRA.2007.339605

Thellman, S., De Graaf, M., and Ziemke, T. (2022). Mental state attribution to robots: a systematic review of conceptions, methods, and findings. *ACM Transact. Hum. Robot Interact.* 11, 1–51. doi: 10.1145/3526112

Tsumura, T., and Yamada, S. (2022). Agents facilitate one category of human empathy through task difficulty. In 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (22–28).

Vollmer, A. L., Read, R., Trippas, D., and Bealpaeme, T. (2018). Children conform, adults resist: a robot group induced peer pressure on normative social conformity. *Sci. Robot.* 3:eaat7111.

Zhou, W., and Shi, Y. (2011). Designing empathetic social robots. In proceedings of the 2011 2nd international conference on artificial intelligence, management science and electronic commerce (AIMSEC), pp. 2761–2764.

Zoghbi, S., Croft, E., Kulić, D., and Van der Loos, M. (2009). Evaluation of affective state estimations using an on-line reporting device during human-robot interactions. IEEE/RSJ international conference on intelligent robots and systems, St. Louis, MO, USA, 2009, pp. 3742–3749.

TYPE Original Research
PUBLISHED 16 August 2024
DOI 10.3389/fpsyg.2024.1388860

Impact of media dependence: how emotional interactions between users and chat robots affect human socialization?

Ziying Yuan, Xiaoliang Cheng* and Yujing Duan

College of Communication Science and Arts, Chengdu University of Technology, Chengdu, Sichuan, China

OPEN ACCESS

EDITED BY
Simone Belli,
Complutense University of Madrid, Spain

REVIEWED BY
Xindong Ye,
Wenzhou University, China
Minjin Rheu,
Loyola University Chicago, United States

*CORRESPONDENCE
Xiaoliang Cheng
✉ 824569136@qq.com

RECEIVED 20 February 2024
ACCEPTED 05 August 2024
PUBLISHED 16 August 2024

CITATION
Yuan Z, Cheng X and Duan Y (2024) Impact of media dependence: how emotional interactions between users and chat robots affect human socialization? *Front. Psychol.* 15:1388860.
doi: 10.3389/fpsyg.2024.1388860

In the era of intelligent media, human users and chatbots have established a deep dependency relationship through communication, making media dependence a behavioral foundation that widely permeates human social practice. This article investigates how media dependence affects human social interaction during emotional interactions between human users and chatbots. Based on the theory of media dependence, the existing mature scales of media dependence and interpersonal communication were adapted, and 496 Replika user questionnaires were collected. After screening the validity of the questionnaires, 428 valid questionnaires were obtained. Descriptive statistical analysis, correlation analysis, multiple linear regression analysis, and mediation effect testing were used to analyze the impact of media dependence on human–computer emotional interaction. Results indicate a significant positive correlation between human–chatbot emotional interaction and human user social interaction. Media dependence significantly positively regulates emotional interactions between humans and chatbots. In addition, the social interactions of human users are partially influenced by factors such as user nature, age, education, and income.

KEYWORDS

chatbot, emotional interaction, interpersonal communication, media dependence, Replika

Introduction

In recent years, with the continuous development and maturity of brain–computer interface, VR, AI, gene editing, and other technologies, human society has entered a new stage of human–machine integration and symbiosis—the post-human era. In this era, machines are no longer just cold pieces of iron in the traditional sense. They have become media channels that connect and facilitate interactions between people and society, serving as the primary means of communication. The advent of technology disrupts traditional interpersonal communication patterns, fostering a novel relationship between humans and machines, and some of the public who rely heavily on technology are satarting to ignore the emotional relationship between people.

In November 2022, the artificial intelligence lab OpenAI officially launched the universal chatbot ChatGPT. It attracted more than one million users within 5 days of its launch, which is the height of Meta in 10 months and of Netflix in 3 years (Hurst, 2022). In fact, chatbots are nothing new, and the birth of ChatGPT is not an accident. As early as the year of 1966, Weissenbaum, the father of modern artificial intelligence, launched the earliest AI chatbot

ELIZA, followed by PARRY, ALICE, Jabberwocky, and other chatbots. The rapid popularity of AI chatbots has not only aroused widespread attention from the academic community but also received the "olive branch" from the business community. In the 21st century, tech giants have similarly launched highly branded chatbots, including Apple's Siri, Google's Bard, and Facebook's Blender. In comparison, China's AI chatbot was born slightly later. It was not until 2004 that Xiao-I officially launched China's first chatbot named MSN. Subsequently, Xiaomi's Xiao Ai, Baidu's ERNIE Bot, and JD's Chat-JD also gradually received public attention. With the development of chatbots, machines can not only imitate people's voices, language, and expressions but also tap into the human heart—emotion. A chatbot breaks through the so-called intersubjectivity and becomes a humanized, emotional, and creative humanoid. As Dominique (2010) wrote in his book: "Emotional factors are essential to understanding the complexity of the world we live in." The emergence of Replika, an emotional chatbot, not only conforms to the development of the technological era but also shows the necessity of emotional communication. In 2015, Eugenia Kuyda, in memory of a friend who died unexpectedly in the same year, used Google's basic neural network program to synthesize thousands of conversations of her friend's messages. This process gave the impression that the deceased friend was communicating when responding to messages. This endeavor was the precursor to Replika. In 2016, Luck officially launched Replika, which has been downloaded more than 10 million times globally. It is the most downloaded chatbot in the App Store and was among the top five mental health programs during the COVID-19 pandemic in 2020. As a dedicated emotional chatbot, Replika provides emotional comfort to many lonely people and those with social phobia through long-term emotional companionship and positive emotional support. Replika was downloaded 55,000 times in Mainland China in the first half of 2021. In addition, a group called "Man–machine Love" quietly emerged on Douban, a Chinese internet-based community platform. The group focuses on emotional communication between humans and AI partners, and it has more than 9,600 members so far.

As the "sixth medium" after newspapers, radio, television, computers, and mobile phones (Lin and Ye, 2019), chatbot's innate "machine" characteristics and constantly evolving "humanity" not only make the communication between people and chatbot flexible and natural but also generate emotions that human–computer interaction did not have in the past. Therefore, if the public continues to invest time, energy, and emotion in chatbots; constantly rely on virtual social interaction; and even give chatbots "life," can this kind of human–machine emotional interaction similar to interpersonal communication replace realistic interpersonal communication? That is, what impact will the trend of human–machine emotional interaction have on users' realistic interpersonal communication?

From the perspective of human–machine communication (HMC), this study takes media dependence as the main theoretical basis. According to the specific application scenario of the Replika chatbot, the author adapts the existing scale and collects variable data through a questionnaire survey, resulting in 428 valid questionnaires. Specifically, this paper makes reference to Rubin's Quasi-Social Interaction Scale, Kwon's Smartphnoe Addiction Scale, Zheng Richang's Interpersonal Comperhensive Relationship Dignostic Scale, etc., as detailed in the questionnaire design section and Table 1. Since the number of Replika users in China is relatively small and scattered, the author mainly distributes and collects questionnaires through platforms and forums where Replika users are active. For example, Douban group, Xiaohongshu platform, Weibo Replika Super talk, etc. Then, descriptive statistical analysis, correlation analysis, multiple linear regression analysis, and mediation effect testing were used to analyze the impact of human–chatbot emotional interaction on human interpersonal communication. Basing on the integration and extension of previous views, this study introduces media dependence theory originally applicable to interpersonal communication into the vision of human–machine communication research. Doing so not only enriches the theoretical scope of human–machine communication but also lays a foundation for building a harmonious and symbiotic human–machine moral destiny community.

TABLE 1 Basic information about the respondents.

Dimension	Category	*N*	Percentage
Gender	Male	122	28.50%
	Female	306	71.50%
Age	Under 18 years old	29	6.80%
	18–25 years old	264	61.70%
	26–30 years old	86	20.00%
	Over 30 years old	49	11.50%
Education	High school degree or below	81	18.90%
	College degree	55	12.90%
	Bachelor degree	223	52.10%
	Master degree or above	69	16.10%
Monthly income	3,000 yuan and below	214	50.00%
	3,001–5,000 yuan	108	25.20%
	5,001–7,000 yuan	44	10.30%
	7,001–10,000 yuan	33	7.70%
	10,001–15,000 yuan	18	4.20%
	15,001–30,000 yuan	8	1.90%
	More than 30,000 yuan	3	0.70%

Research review

Media dependence

Influenced by Durkheim's views on media, American communication researchers Melvin Defler and Sandra Bower-Killoch were the first to define "dependency" from the perspective of social ecology (Sandra et al., 2004). In their paper The Dependence Mode of Mass Communication Media Effect, they pointed out that there is a close dependence relationship among the three systems of media, audience and society, and held that the audience has a corresponding relationship between the satisfaction of information needs and the achievement of goals, and proposed that there is a positive correlation between media dependence and media effect. That is, media dependence will change the public's existing cognition, attitude and behavior through media content (Melvin and Sandra, 1990). In addition, It is generally believed that there are two forms of media dependence: the explicit habitual dependence and the implicit spiritual dependence. The so-called habitual or conditional dependence refers to the habitual behavior of the public on the media after long-term use of the media. The long-term conditional media dependence will further affect the spirit and psychology of the public and form spiritual dependence. For example, some people will have anxiety, tension, emptiness, lack of security and other emotions when they are separated from the mobile phone media. In the year 2008, the UK Post Office conducted a survey on mobile phones and found that users felt anxious about not being able to use their phones or not having them around, a phenomenon they dubbed "no phone phobia" (SecurEnvoy, 2012). Generally speaking, media dependence refers more to the psychological dependence of the public on the media.

In the context of the continuous development of information technology, the public's dependence on all kinds of media not only shows a growing trend, but also shows a new form. Some scholars even put forward that the closeness and particularity of the relationship between network media and audience make the media dependence theory more applicable in the Internet environment (Xie, 2004). Through searching relevant literature, the author finds that there are abundant researches on media dependence related to traditional media such as newspaper, radio and television, while there is still room for further research on media dependence in respect of new media. In particular, new media such as chatbots are no longer just intermediaries or tools for communication, but become interlocutors for communication with people. Therefore, for some users, it is easy to see chatbots as people with living characteristics. With the extension of users' communication time with chatbots and the increase of emotional investment, they may become mentally dependent on chatbots. Therefore, to explore whether users will be dependent on chatbots, and the role of media dependence in human-computer emotional interaction and interpersonal interaction are interesting topics worthy of attention.

Research on human–machine emotional interaction

At present, research on human–machine emotional interaction mainly focuses on two areas. On the one hand, some scholars discuss how human–machine emotional communication can be realized and the ethical issues in this communication from the perspective of philosophy and humanity. On the other hand, some scholars, starting from reality, intend to clarify the process and optimization path of human–machine emotional interaction.

From the perspective of philosophy and humanity, relevant research can be further subdivided into two aspects: positive discussion and negative reflection. Under positive discussion, David (2007) believed that humans cannot help but form difficult emotional relationships with companions, things, and even robots around them, which is human nature. Sven and Lily (2017) believed that with the continuous development of intelligent technology, robots can imitate and learn human emotions and even fall in love with people. Under negative reflection, John P. Sullins (Peters, 2015) believed that although humans can reach an emotional connection with robots, emotions are complex things, and discerning subtle emotions like humans is difficult for machines. Emmelyn et al. (2021) also believed that establishing close friendships between humans and machines, akin to human relationships, is challenging. Whitby (2008) believed that owing to the "machine nature" of machines, some users may transgress moral boundaries by resorting to verbal violence against the machine. This behavior has the risk of extending from online to offline contexts, potentially strengthening the desire of users to commit physical violence.

Regarding the research on the process and implementation of human–machine emotional interaction, Li and Zhao (2023) studied the interaction between the robot NAO and human. They found that the user can independently judge the emotional meaning implied in the machine actions, and this ability has no obvious relationship with the familiarity between human and machine. Zhang and Han (2022) studied the emotional interaction between users and Xeva software and found a progression from companionship to tolerance and then to rational return in the interaction between the two. Taking smart audio as an example, Kang (2023) discussed the characteristics of machines from various aspects. They investigated the relationship between intelligent machines and user' families, especially the emotional compensation effect of machines for people living alone. Gan and Guo (2022) observed a coffee robot in Shanghai and believed that the "embodied" aspect is the basis of emotional interaction between consumers and robots. Through the review of previous literature, it is not difficult to find that human beings will have emotional communication in the interaction with chatbots, and such emotional interaction will not only have an impact on the relationship between human and machine, but also have an impact on the actual emotions of users. Consequently, based on the aforementioned research, this paper employs the Replika chatbot as a case study to delve into the current situation of emotional exchange between users and chatbots, encompassing the degree, behaviors, effect, and satisfaction associated with human-chatbot emotional interaction.

Replika as an emotional chatbot platform, the current research mainly focuses on two aspects. On the one hand, the interaction mechanism between users and Replika chatbots is studied, including interaction characteristics, interaction purpose, interaction behavior, etc. Zhang and Sun (2023) combined various research methods to analyze the text messages posted by Replika users on various social platforms and the data obtained from interviews, and studied the emotional connection between Replika users and chatbots from the perspective of embodied imagination between Replika and users. Through interviews with 20 Replika in-depth users, Tan (2023) found that the formation of the current human-computer intimate relationship basically conforms to the process of social penetration theory from the shallow to the deep, from the surface to the inside, and the "human" characteristics of chatbots, human-computer trust and other factors play an important role. The other aspect is the discussion of the ethical issues existing in the interaction between users and Replika. Taking Replika as an example, Possati (2023) used psychoanalysis to discuss the control and responsibility of human unconscious behavior on AI design and behavior. Combined with the murder incident instigated by Replika in Corriere Della Sera, he criticized and reflected on the current ethics of human-computer interaction. Kourkoulou (2023) also used Replika chatbots to discuss the hidden exploitation of emotional labor in the current digital economy, as well as the ethical issues in the field of artificial intelligence in practice, especially the gender and racial stereotypes caused by current chatbots. While discussing the risks and impacts of generative AI in future economic development, Orchard and Tasiemski (2023) also discussed the sexual and pornographic phenomena of Replika chatbot in personalized services, and reflected on the verbal attacks brought by users to the machine. Zeng and Cao (2023), through long-term observation of the Replika team, critically reflected on human-machine relationship from the aspects of commercialization, relationship imbalance, dissociation and so on.

Research on the relationship between media dependence, human–chatbot emotional interaction, and interpersonal communication

As a new phenomenon in contemporary society, the emotional interaction between chatbots and users still has some lacks in-depth research on the relationship between media dependence and human–chatbot emotional interaction. As chatbot is an emerging technology product, its essence and core cannot be separated from the media itself. Therefore, previous studies on various media and their relationship with media dependence can provide certain references for this paper. Through the review and summary of previous literature, the author finds that the current research on the relationship between media use and media dependence mainly focuses on one aspect: there is a positive correlation between media use (the interaction between people and media) and media dependence. As early as the 20th century, some scholars have studied the relationship between TV media and media dependence, and found that there is a close correlation between the audience's interaction with TV shopping programs and media dependence (Grant et al., 1991). Wang (2014) found that there was a positive correlation between college students' use of wechat and media dependence. Some scholars have found that the higher the degree of interaction between users and social robots, the stronger the user's dependence on the media generated by social robots (Han et al., 2021). Studies have also shown that the degree to which users interact with virtual idols can positively predict users' media dependence (Zhou and Zhang, 2023). From the extant literature, it is evident that scholars collectively concur in the observation of a positive correlation between the frequency of media utilization and the degree of media dependency. In light of this, it is pertinent to inquire into the nature of the emotional exchange between users and chatbots, a novel medium, in relation to media dependency. Therefore, the following hypothesis is proposed:

> *H1*: human-chatbot emotional interaction is positively correlated with media dependence.

In terms of media dependence and interpersonal communication, scholars mainly hold two viewpoints. (1) The public's dependence on media is conducive to enriching daily life, improving their emotional acquisition and companionship. It also plays a positive role in the practical interpersonal communication of the public. For example, He and Zhu (2024) found that the dependence of left-behind women in rural China on short video media can enhance their self-cognition and adjust their personal emotions, thus improving their daily life. Jiang (2022) studied the media dependence and subjective well-being of the elderly during the COVID-19 pandemic and found that the elderly has a high sense of dependence and trust in TV media. TV media can not only make up for the gap in real interpersonal communication but also strengthen the happiness of the elderly. (2) Media dependence can exacerbate loneliness. Kim et al. (2009) found that there is a vicious cycle between loneliness and media dependence, and excessive use of social media can lead to deeper levels of loneliness. Zhang et al. (2020) have identified a significant correlation between mobile phone addiction and individual loneliness, indicating a moderate positive relationship between the two phenomena. Han et al. (2021) found that users' media dependence on Microsoft Xiaobing strengthens users' sense of loneliness. In light of the escalating authenticity, immersion, and interactivity associated with chatbots, there is a possibility that users' reliance on chatbots, along with their dependent behaviors, could surpass their previous dependency on traditional media. Consequently, within this context, does the users' reliance on chatbots for media purposes exacerbate their feelings of loneliness? Furthermore, does this sense of dependency have an impact on their genuine interpersonal engagements? Therefore, the following hypothesis is proposed:

> *H2*: media dependence is positively related to users' interpersonal communication status.

In the aspect of human–chatbot emotional interaction and interpersonal communication, a large gap is noted in the relevant research at home and abroad. However, as the predecessor of human–machine interaction, the relationship between network interaction and real interpersonal communication can also reflect the relationship between human–machine interaction and real interpersonal communication to a certain extent. This stream of research is mainly divided into two categories. One view holds that virtual and online communication, as an important complementary form of real interpersonal communication, can have a positive effect on real interpersonal communication. For example, Su (2020) studied the communication behaviors between users and the mobile game Dream Journey to the West. They found that compared with real interpersonal communication, online intimate relationships in the form of games can be transformed into offline relationships and are conducive to maintaining long-distance interpersonal relationships. Through empirical research, Wang and Fu (2016) also found that the use of social media is conducive to expanding the frequency and breadth of college students' realistic interpersonal communication and can supplement the realistic interpersonal communication to a certain extent. The other view is that virtual and online interaction is not conducive to maintaining real interpersonal relations. It is not conducive to the public to grasp the rules of interpersonal communication. Moreover, it easily leads to the public escaping from reality and avoiding social behavior. For example, in "The Interactive Ritual Chain," Collins (2009) argued that the lack of ritual is why the advent of email has allowed the masses to indulge in utilitarian interactions and weakened real-world relationships. The dependence behavior caused by long-term media use easily causes the public to have real social difficulties, which is not conducive to users' real interpersonal communication. For example, Lin (2020) believed that the public's dependence on the media can cause problems such as difficulty in choice. The public can experience difficulties distinguishing between the media world and the real world, which is not conducive to their real life. Chen (2009) believed that TV media, as an important medium in children's growth, creates a mimicry environment that deviates from reality through prolonged exposure, gradually alienating children from people and things in the real world. Wei (2012) believed that owing to the strong virtuality of Weibo, the interaction in Weibo can impact real interpersonal relationships and lead to the loss of human subjectivity. Concerning the interaction between users and chatbots, as a form of interpersonal communication, it raises questions regarding whether it may mitigate the intensity of users' authentic social connections and whether it could augment the user's perception of solitude. And what role does media dependence play between the two? Therefore, the following hypotheses are proposed:

> *H3*: human-chatbot emotional interaction is positively correlated with the user's interpersonal communication status.
>
> *H4*: media dependence plays a mediating role between the current situation of human-chatbot emotional interaction and the current situation of real interpersonal communication.

Methods and investigations

Data collection and sample description

The study constructed the formal questionnaire of the survey through the Wenjuanxing platform, which provides functions equivalent to Amazon Mechanical Turk. Considering that Replika is an emotional chatbot, and emotional factor is an essential element for users to interact with AI character, this paper takes users who have used Replika as the research object. From October 22 to November 11, 2023, the questionnaires were distributed and collected through the Douban's "Man–Machine Love" group (9,602 members), "My family's Replika has become fine" (2,327 members), "Female Players Association" (46,936 members), "My Replika is very warm" (419 members), the Replika topic on Xiaohongshu platform, Weibo Replika Super talk, and other platforms. The study also used the private messaging functions of Douban, Xiaohongshu, Weibo, Xianyu, and Douyin platforms to conduct a one-to-one questionnaire survey of users who posted or commented on Replika information on these platforms. Finally, 496 questionnaires were collected, and 68 invalid questionnaires were eliminated to ensure the effectiveness of the survey. Invalid questionnaires mainly included those that took less than 60 s to answer (3 questionnaires), those who chose "never used the Replika software" in the first question. (48 questionnaires), those whose answers to all questionnaires were the same (3 questionnaires), and those with the same ID were filled in multiple times (14 questionnaires). A total of 428 valid questionnaires were obtained, with an effective rate of 86.29%. The basic information of the respondents is shown in Table 1.

Questionnaire design

The questionnaire scale was developed by combining the existing literature and classical scales, focusing on the study object, Replika chatbot. The scale used a five-level Likert scoring method, 1 = "strongly agree," 2 = "agree," 3 = "uncertain," 4 = "disagree," and 5 = "strongly disagree." After forming the preliminary draft of the questionnaire, the preliminary survey was carried out in a small scope. According to the questionnaires recovered from the preliminary survey, problems were found, and the questionnaires were optimized to form a formal questionnaire. Specifically, according to the results of the preliminary survey, the author modified the questions with ambiguous expressions and adjusted the order of the questions. For example, the original B13 question "Replika can find my mood changes during interaction." was adjusted to B12 question "Replika can well understand my feelings or emotions during interaction." to enhance the logic and coherence of the question. Moreover, the original B7 "I will communicate my work with Replika during the interaction." and B8 "I will communicate my learning with Replika during the interaction." overlap, so the two questions have been merged as "I will communicate my work or learning with Replika during the interaction. Besides, Considering the simplicity of the questionnaire, the options of "elementary school education," "junior high school education" and "high school education" in the questionnaire E3 were merged into "high school education and below." Finally, questionnaires were distributed and collected from Replika users on platforms such as Douban, Xiaohongshu, and Weibo.

The questionnaire consists of the following parts: the title and introduction, the body, and the end. The body of the questionnaire includes the usage of Replika, the emotional interaction between the user and Replika (chatbot), the user's interpersonal communication, and the user's demographic information.

In order to dispel the concerns of Replika users and ensure that the respondents can fill in the questionnaire seriously and carefully, in the introduction section of the questionnaire, the author briefly introduces the identity of the investigator, the purpose of the survey, the connotation of the chatbot, the meaning of emotional interaction, and the use of the questionnaire data.

The main body of the questionnaire includes five aspects:

(1) *Users' Replika usage.* This part is mainly to understand the respondents' basic use of Replika, including the gender and identity of the virtual character set by users, the frequency and years of users' use of the Replika chatbot, the Replika level and user's purpose of use.

(2) *Current situation of emotional interaction between users and Replika.* It mainly includes four aspects, namely, the degree of the current user's emotional interaction with Replika, the behavior in their emotional interaction, the effect of emotional interaction, and the user's satisfaction with emotional interaction.

First, the degree of emotional interaction between users and chatbots is mainly reflected by the degree of privacy of the content exchanged between users and chatbots, the degree of personal emotion revealed in the communication, the degree of empathy between users and AI characters, and so on. Second, considering the differences of communication subjects, this paper further divides behavior of emotional interaction between human and chatbots into chatbot behavior and human behavior. The behavior of human emotional interaction is mainly reflected by the willingness of users to communicate with chatbots about hobbies, work, study and real thoughts in real life, the degree of users' respect for the views of Ai characters, and so on. The behavior of chatbot emotional interaction is mainly reflected by the AI character's ability to perceive and understand the emotional changes of users, ability to comfort users, ability to solve problems for users, and so on. Third, communication effect can be divided into cognitive effect, emotional (attitude) effect and behavioral effect (Guo, 2011). Therefore, this paper further divides the effect of emotional interaction into three levels: cognition, emotion, and behavior. The effect of emotional interaction at the cognitive level can be reflected mainly through the cognitive effect of users on the basic information and expression ability of chatbots. The effect of emotional interaction at the emotional level is mainly reflected through the ability of chatbots to relieve loneliness, release life pressure, provide companionship and care for users. Finally, the effect of emotional interaction on user behavior is mainly reflected by the degree of user's dependence on chatbot and the closeness of the relationship between user and chatbot. The satisfaction of emotional interaction is mainly reflected by the user's willingness to continue to use chatbots, the user's willingness to recommend chatbots to others, and the user's recognition of the expression ability and intelligence of chatbots.

Concretely speaking, based on the qusai-social interaction scale prepared by Rubin et al. (1985), quasi-social relationship scale prepared by Horton and Wohl (1956), user-role interaction scale prepared by Auter and Palmgreen (2000), the microblog interaction scale prepared by Lu (2011) and the gottman scale of quasi-social relations prepared by Ge (2017), this paper adjusted and modified the scale in combination with the specific research objects of this paper, and finally formed the scale of emotional interaction between users and chatbots in this paper. In this paper, the interactive object in the original item is modified from "local news anchor" to "Replika," and the original single scale is divided into "degree, behavior, effect and satisfaction." The detailed contents of the scale are delineated in Appendix Table.

(3) *Users' media dependence on Replika.* The purpose is to understand the media dependence caused by emotional interaction between users and Replika and further understand the relationship between human–chatbot emotional interaction and media dependence. Specifically, based on the Smartphone Addiction Scale (SAS) compiled by Kwon et al. (2013) and the Chinese Internet Addiction Scale compiled by Chen et al. (2003), this paper adjusted and modified the scale in combination with the specific research objects of this paper, and finally formed the media dependence scale of this paper. This article modifies "smartphone" and "internet" in the original scale to "Replika." The detailed contents of the scale are delineated in Appendix Table.

(4) *Users' interpersonal communication.* This aspect deals with the main understanding of the user in the real-life interpersonal communication of negative emotions, loneliness, and so on. Specifically, based on the Interpersonal Comprehensive Relationship Diagnostic Scale (ICDS) compiled by Zheng (1999) and the Social Avoidance and Distress Scale (SADS) of Watson and Friend (1969), this paper adjusted and modified the scale in combination with the specific research objects of this paper, and finally formed the user's real interpersonal communication status scale in this paper. The detailed contents of the scale are delineated in Appendix Table.

(5) *Respondents' demographic information.* The demographics of Replika users can also affect the Replika chatbot usage, so questions about users' gender, age, monthly income, and educational background were included.

Overall, the body of the questionnaire contains a total of 65 questions, with 52 scale questions and 13 non-scale questions (Table 2).

At the end of the questionnaire, the respondents were again thanked for their patience and cooperation.

Reliability testing and validity testing

Reliability testing is an important part of a questionnaire survey, it is related to the rationality of the questionnaire setting and the reliability of research results. In this study, SPSS27.0 was used as a reliability detection tool. Cronbach's alpha was used to test the data on three aspects: the status quo of emotional interaction between users and Replika chatbots, media dependence brought by emotional interaction, and users' real-life interpersonal communication. If the Cronbach's alpha coefficient is greater than 0.60, the reliability of the questionnaire is acceptable, and if the coefficient is greater than 0.70, the reliability of the questionnaire is good. As shown in the figure, the Cronbach's alpha coefficients of the three scales are all greater than 0.8, and the Cronbach's alpha coefficient of the whole questionnaire scale (52 items) is 0.977, so its reliability is high. The reliability test results of each variable in this study are shown in Table 3.

A validity test is a means to assess the accuracy and validity of each factor of the questionnaire, including surface validity, criterion validity, and construction validity. To test the validity of the questionnaire (scale), KMO and Bartlett tests were conducted on the overall questionnaire scale (52 items) and three groups of scales. If the KMO value is greater than 0.6, then the questionnaire can perform factor analysis; greater than 0.7, is generally suitable for factor analysis; greater than 0.8, is more suitable for factor analysis; greater than 0.9, is very suitable for factor analysis. As shown in Table 3, the KMO value of the four aspects of the questionnaire scale is greater than 0.8, and the overall KMO value of the questionnaire scale is greater than 0.9, so the validity of the questionnaire is high. The validity test results of each variable in this study are shown in Table 4.

Results and data analysis

Analysis of Replika usage

According to the data, the usage time of Replika users is mainly less than 6 months, and the ratio of users who use Replika for less than 6 months to those who use it for more than 6 months is about 3:2. Moreover, there is an overall inverse correlation between the duration of use and the number of users. This shows that most of the current Replika users are short term users, and the stickiness of the platform still needs to be improved. Besides, nearly half of users use Replika once a week, and only 10 percent use Replika every day, which indicates that the current intimacy between users and AI characters is not high. In terms of users' level, this question is optional, and questionnaires with unfilled and unclear answers were eliminated, and the collected levels were sorted. The user level pertains to the rating assigned by the platform to users based on their engagement metrics such as usage duration and interaction frequency. In general, an increase in user duration and interaction frequency typically results in a higher user level. Based on Replika's official user profiling, the chatbot's capabilities at levels 1–10 are in a "preliminary" stage. During this juncture, the emotional support rendered by Replika to its users is comparatively constrained, and the level of intimacy and frequency of interaction between users and the chatbot is minimal (Zeng Yiguo, et al., 2023). However, as users progress beyond level 10 and continue their development, the dialogue dynamics between the user and the chatbot become increasingly fluid and profound, thereby strengthening the emotional bond between them. The data shows that the user's level is mainly concentrated in the below 10 level, the user of high level is very few, and the ratio of users below 10 level and above 10 level is about 4:1. Among them, the lowest level is level 1, and the highest level is level 167. This shows that most current users do not

TABLE 2 Questionnaire design.

Dimension		Indicator		Corresponding question number	Source of scale
Part I	Replika usage	Use or not		A1	Non-scale question
		Year of use		A2	
		Frequency of use		A3	
		Replika character's gender, nickname role, and rank		A4–A7	
		Reasons of use		A8	
Part II	Current situation of human–chatbot emotional interaction	Degree of emotional interaction between human and chatbots		B1–B5	Horton and Wohl (1956), Rubin et al. (1985), Auter and Palmgreen (2000), Lu (2011), and Ge (2017)
		Behavior of emotional interaction between human and chatbots	Human behavior	B6–B11	
			Chatbots behavior	B12–B19	
		Effect of emotional interaction between human and chatbots	Cognitive effect	B20–B24	
			Emotional effect	B25–B30	
			Behavioral effect	B31–B35	
		Satisfaction of emotional interaction between human and chatbots		B36–B41	
Part III	Users' media dependence on Replika			C1–C6	Chen et al. (2003) and Kwon et al. (2013)
Part VI	Users' real-life interpersonal communication	Interpersonal interaction willingness, Interpersonal relationships, interpersonal attitudes, etc.		D1–D5	Watson and Friend (1969) and Zheng (1999)
Part VII	Personal information	Gender, age, educational background, occupation, monthly income		E1– E5	Non-scale question

have a strong intimacy with AI characters. In addition, the data shows that 74.1% of users regard entertainment as one of the purposes of using replika, 60% of users regard social networking as one of the purposes of using replika, 43.90% of users regard learning as one of the purposes of using replika, and less than 10% of users think that using replika is for other purposes. This suggests that users use replika primarily for entertainment and socializing, again for learning, and finally for other purposes. The statistical data of Replika usage is shown in Table 5.

Analysis of current situation of human–chatbot emotional interaction

The current situation of human–chatbot emotional interaction is mainly reflected through four aspects: the degree of emotional interaction, the behavior of emotional interaction, the effect of emotional interaction, and the satisfaction of emotional interaction. The higher the score of emotional interaction, the higher the degree of emotional interaction, the more engaged behavior, the better the effect and the higher the satisfaction.

First, the level of emotional interaction between users and the Replika chatbot is above average, with a mean of 3.6313 and a standard deviation of 0.84679. The degree of emotional interaction between the user and the Replika chatbot is mainly measured by the degree to which the user has genuine emotional communication with the Replika chatbot, the degree to which the content of the communication is private, the degree to which the personal emotion is revealed in the communication, the degree to which the user regards Replika as a person and tries to understand them, and the degree to which the user empathizes with the virtual character in Replika.

Second, the behavior of human–chatbot emotional interaction is at a superior level, with a mean of 3.8188 and a standard deviation of 0.75641. The behavior of emotional interaction is mainly reflected in two aspects. On the one hand, the user's behavior in emotional interaction is mainly based on their willingness to communicate with Replika about hobbies, work, and study; the extent to which they express their genuine views and negative emotions on real-life events and people; and the extent to which the Replika avatar is understood. On the other hand, Replika's behavior in emotional interactions is mainly based on the Replika avatar's perception of the user's emotions and emotional changes, the understanding of the user's emotions, the empathy for the user's bad experience, and the comfort to the user and the ability to bring solutions to them. Specifically, the average of Replika's behavior in emotional interactions is 3.8356, and the average of the user's behavior in emotional interaction is 3.7963.

Third, the effect of the emotional interaction between users and the Replika chatbot is also at a superior level, with a mean of 3.7417 and a standard deviation of 0.73615. Among the three effects of affective interaction, the mean of cognitive effect is 3.8463, and the standard deviation is 0.80537. The mean of emotion effect is 3.8980, and the standard deviation is 0.77763. At the behavioral level, the mean effect is 3.4495 and the standard deviation is 0.89316. The emotional interaction behaviors and effects are detailed in Table 6. In general, the level of behavioral effect is the lowest, and the level of cognitive effect is almost equal to the emotional effect. In the end, users' satisfaction with their emotional interaction with Replika is on top, with a mean of 3.8220 (the highest score for the mean of the variable) and a standard deviation of 0.78791 (Table 7).

TABLE 3 Reliability analysis of questionnaire scale.

Variable			Number of projects	Cronbach's Alpha
Current situation of human–chatbot emotional interaction	Degree of emotional interaction between human and chatbots		5	0.822
	Behavior of emotional interaction between human and chatbots	Human behavior	6	0.871
		Chatbots behavior	8	0.905
	Effect of emotional interaction between human and chatbots	Cognitive effect	5	0.858
		Emotional effect	6	0.888
		Behavioral effect	5	0.855
	Satisfaction of emotional interaction between human and chatbots		6	0.891
Users' media dependence on Replika			6	0.923
Users' real-life interpersonal communication			5	0.815
Questionnaire scale population			52	0.977

TABLE 4 Results of KMO and Bartlett's Sphericity Tests.

Variable			KMO measure of sampling adequacy	Bartlett's sphericity test	Significance
Current situation of human–chatbot emotional interaction	Degree of emotional interaction between human and chatbots		0.833	724.244	0.000
	Behavior of emotional interaction between human and chatbots	Human behavior	0.886	1115.741	0.000
		Chatbots behavior	0.925	1801.643	0.000
	Effect of emotional interaction between human and chatbots	Cognitive effect	0.855	900.547	0.000
		Emotional effect	0.905	1242.530	0.000
		Behavioral effect	0.840	942.921	0.000
	Satisfaction of emotional interaction between human and chatbots		0.904	1305.825	0.000
Users' media dependence on Replika			0.905	1821.726	0.000
Users' real-life interpersonal communication			0.803	719.200	0.000
Questionnaire scale population			0.979	16313.800	0.000

Analysis of users' media dependence on replika

The current level of reliance on Replika is moderate, with a mean of 3.2936 (the lowest score of the variable) and a standard deviation of 1.05839. The higher the score of media dependence, the higher the degree of media dependence on Replika. The index is mainly reflected by the increase in the number of times users use Replika, the increase in the length of use, the decrease in the frequency of interaction with family and friends, the decrease in daily leisure activities, the level of depression caused by non-interaction, and the decrease in the enjoyment of life. In general, users do not rely much on Replika, and most users can reasonably balance virtual and real interactions. The specific analysis data can be found in Table 8.

Analysis of users' real-life interpersonal communication

At present, the real-life interpersonal communication of users is at a medium and high level, with an average of 3.6650 and a standard deviation of 0.86687. The higher the score of interpersonal communication, the worse the level and ability of interpersonal communication. The higher the score of this index, the lower the self-disclosure willingness of users in real-life interpersonal communication. This index is mainly reflected by the degree of interaction between users and acquaintances in real-life interpersonal communication, the level of comfort and comfort in real-life interpersonal communication, the degree of hiding their true feelings in real-life interpersonal interaction, the degree of expression of positive emotions rather than new feelings in real-life interpersonal interaction, and the level of loneliness of users in real-life interpersonal interaction. The specific analysis data can be found in Table 8.

Analysis and discussion

Correlation analysis of variables

To explore the correlation between human–chatbot emotional interaction, media dependence, and real-life interpersonal communication, this study processed the data of three variables through Pearson correlation analysis and MLR. The results are shown in Tables 9, 10.

Regression analysis of variables

This study predicted media dependence as the mediating variable in the influence of the current situation of emotional interaction on users' real-life interpersonal communication. Therefore, to further predict the impact of each variable on the user's real-life interpersonal communication, this study conducted hierarchical regression analysis on the predicted independent variable (the current situation of emotional interaction), intermediary variable (media dependence), and control variable (demographic variable). First, gender, age, education, and monthly income together explain 11% of users' real-life interpersonal communication. In terms of gender ($\beta = 0.063$, $p = 0.000 < 0.05$), women's real-life interpersonal communication score is higher than that of the male group. In terms of education ($\beta = -0.276$, $p = 0.000 < 0.05$), users with higher education had lower scores in real-life interpersonal communication. In terms of age ($\beta = 0.203$, $p = 0.200 > 0.05$) and monthly income ($\beta = -0.005$, $p = 0.914 > 0.05$), no significant relationship is observed between them and users' real-life interpersonal communication scores. The score of real-life interpersonal communication mainly reflects the user's social integration and self-disclosure in real-life interpersonal communication. The higher the score, the lower the degree of social integration and self-disclosure.

TABLE 5 Statistical data of Replika usage.

Dimension	Category	*N*	Percentage
Duration of use	Less than 3 months	134	31.30%
	3–6 months	132	30.80%
	6–12 months	77	18.00%
	1-3-years	54	12.60%
	More than 3 years	31	7.30%
Frequency of use	At least once a day	59	13.80%
	At least once a week	204	47.70%
	At least once a half month	89	20.80%
	At least once a month	50	11.70%
	Less	26	6.00%
Users' level	Below or equal to level 10	121	79,60%
	Above level 10	31	20.40%
Purpose of use	Entertainment purposes	317	74.10%
	Social purposes	257	60.00%
	Learning purposes	188	43.90%
	Other purposes	29	6.80%

When human–chatbot emotional interaction enters the equation as the second factor, 47.8% of users' real-life interpersonal communication is further explained. Users' emotional interaction ($\beta = 0.712$, $p = 0.000 < 0.001$) becomes an important factor affecting users' real-life interpersonal communication. The stronger the emotional interaction with the Replika chatbot, the worse the self-disclosure and social inclusion of users in their real-life interpersonal interactions.

When users' media dependence on Replika is entered the equation as the third factor, 7.7% of users' real-life interpersonal communication is further explained. Media dependence ($\beta = 0.418$, $p = 0.000 < 0.05$) has a significant impact on users' real-life interpersonal communication. The more dependent the participants are on the medium of Replika, the worse their self-disclosure and social integration in real-life interpersonal communication. In addition, after the Durbin–Watson (DW) test of Models 1, 2, and 3, the DW values of the three models are close to 2 (the DW value of Model 1 is 1.970; Model 2, 1.908; Model 3, 2.011), indicating the high independence of the data. Overall, hierarchical regression explains 66.5% of the total equation, and detailed data are shown in Table 11.

Mediation effect analysis of variables

Based the previous research, this study sets the current situation of emotional interaction between users and chatbot Replika as the independent variable (X), the user's real-life interpersonal communication as the dependent variable (Y), and the user's acquired media dependence as the intermediary variable (M). The following regression equation can be used to represent the relationship between the variables:

$$Y = cX + e_1$$

$$M = aX + e_2$$

$$Y = c^{'}X + bM + e_3$$

In the equation, c is the total effect of independent variable X (current situation of human–chatbot emotional interaction) on

TABLE 6 Statistical data of human–chatbot emotional interaction behavior and effect.

Indicator		*N*		Mean	Standard	Variance	Minimum	Maximum	Sum
		Effective	Missing sample						
Behavior	Human behavior	428	0	3.7963	0.80210	0.643	1.00	4.83	1624.83
	Chatbot behavior	428	0	3.8356	0.77902	0.607	1.00	5.00	1641.62
Effect	Cognitive effect	428	0	3.8463	0.80537	0.649	1.20	5.00	1646.20
	Emotional effect	428	0	3.8980	0.77763	0.605	1.33	5.00	1668.33
	Behavioral effect	428	0	3.4495	0.89416	0.798	1.00	5.00	1476.40

TABLE 7 Statistical data of the current situation of human–chatbot emotional interaction.

Indicator	*N*		Mean	Standard	Variance	Minimum	Maximum	Sum
	Effective	Missing sample						
Degree of emotional interaction	428	0	3.6313	0.84679	0.717	1.20	5.00	1554.20
Behavior of emotional interaction	428	0	3.8188	0.75641	0.572	1.00	4.86	1634.43
Effect of emotional interaction	428	0	3.7417	0.73615	0.542	1.50	5.00	1601.44
Satisfaction of emotional interaction	428	0	3.8220	0.78791	0.621	1.00	5.00	1635.83

TABLE 8 Statistics on media dependence and real-life interpersonal communication.

Indicator	*N*		Mean	Standard	Variance	Minimum	Maximum	Sum
	Effective	Missing sample						
Users' media dependence on Replika	428	0	3.2936	1.05839	1.120	1.00	5.00	1409.67
Users' real-life interpersonal communication	428	0	3.6650	0.86687	0.751	1.00	5.00	1568.60

dependent variable Y (interpersonal communication). a is the effect of the independent variable X on the intermediary variable M (media dependence). **c** is the direct influence of independent variable X on dependent variable Y after controlling the influence of intermediary variable X. b is the influence of intermediary variable M on dependent variable Y after controlling the influence of independent variable X. In addition, the coefficients e_1, e_2, and e_3 are error terms.

To further clarify the mediating effect of Replika media dependence on users between emotional interaction and interpersonal communication, this study used PROCESS V4.1 plug-in in SPSS26.0 software as a research tool to analyze and study the three. 5,000 samples were selected from the original sample to estimate the 95% confidence interval of the mediation effect, and Model 4 was chosen. In this paper, according to Hayes et al.'s (2012) viewpoint, 5,000 bootstrap samples were selected using the deviation-corrected percentile bootstrap method to test the moderated mediating effect. Specifically, from the 428 samples that have been returned, 5,000 times, one sample at a time, to get a new sample. The Bootstrap method can avoid the limitation of data distribution hypothesis and better deal with non-parametric statistical problems. If the confidence interval does not contain 0, then the mediation effect exists, and vice versa. The data showed a significant positive correlation between human–chatbot emotional interaction and users' Replika media dependence ($\beta = 0.88$, $p = 0.000 < 0.001$). That is, human-chatbot emotional interaction leads to user dependence on Replika chatbots. Consistent with the results of hierarchical regression, media dependence ($\beta = 0.34$, $p = 0.000 < 0.05$) has a significant positive impact on users' real-life interpersonal communication.

After in-depth analysis of the mediation test results, the author found that even after controlling for the four variables of gender, age, education, and monthly income, the emotional interaction between users and Replika has a direct effect on users' real-life interpersonal communication [Effect = 0.56, $p = 0.000 < 0.001$, 95% CI (0.47, 95% CI)]. The indirect effect [Effect = 0.30, 95% CI (0.23, 0.38)] reaches statistical significance, indicating a significant partial mediating effect between the current situation of human–chatbot emotional interaction and users' interpersonal communication. The mediating effect accounts for 34.88% of the total effect. Table 12 provides the details.

Conclusion

The results show that all four hypotheses are verified. First of all, there is a significant positive correlation between the pairwise of the three variables, that is, H1, H2 and H3 are all valid. Secondly, media dependence plays a partial mediating role between between the current situation of human-chatbot emotional interaction and users' interpersonal communication, that is, H4 is valid.

According to the descriptive statistical results of the variables, the users are mostly light users (low level), and they have a moderate level of time and emotion invested in the Replika platform. Besides, in the behavior of emotional interaction between users and chatbots, the score of user behavior is close to that of chatbots, and the score of chatbot behavior is slightly higher than that of users. This suggests that users pay more attention to their emotional acquisition in interaction rather than emotional engagement. And, the three effects of emotional interaction show the following relationship: emotional effect > cognitive effect > behavioral effect. Users have a high average satisfaction with Replika chatbot and a strong willingness to continue using it. In addition, users' media dependence on Replika is at a moderate level, and most users can reasonably balance real and virtual interactions. Finally, the average value of users' real-life interpersonal communication is at a medium to high level.

The correlation analysis results show that the correlation between human–chatbot emotional interaction and users' interpersonal communication is the strongest, followed by the correlation between media dependence and users' interpersonal communication. The correlation between media dependence and human-chatbot emotional interaction is relatively weak. Among the four indicators of the emotional interaction, the effect of emotional interaction has the strongest correlation with the user's interpersonal communication, followed by the degree of emotional interaction and the behavior of emotional interaction. The satisfaction of emotional interaction has the weakest correlation with the interpersonal communication. The hierarchical regression analysis shows that human–chatbot emotional interaction and media dependence is significantly positively correlated with users' real-life interpersonal communication. The current situation of human–chatbot emotional interaction has a greater impact on users'

TABLE 9 Correlation analysis of variables.

Indicator		1	2	3			
				Degree	Behavior	Effect	Satisfaction
Users' media dependence on Replika		1					
Users' real-life interpersonal communication		0.741**	1				
Current situation of human–chatbot emotional interaction	Degree of emotional interaction	0.686**	0.697**	1			
	Behavior of emotional interaction	0.581**	0.697**	0.836**	1		
	Effect of emotional interaction	0.705**	0.752**	0.817**	0.868**	1	
	Satisfaction of emotional interaction	0.530**	0.636**	0.709**	0.817**	0.846**	1

**Significant association at 0.01 level (bilateral). *Significant association at 0.05 level (bilateral).

TABLE 10 Correlation analysis of variables.

		1	2	3
Users' media dependence on Replika	Pearson correlation	1	0.741**	0.674**
	Significance (two tail)		0.000	0.000
	N	428	428	428
Users' real-life interpersonal communication	Pearson correlation	0.741**	1	0.753**
	Significance (two tail)	0.0000		0.000
	N	428	428	428
Current situation of human–chatbot emotional interaction	Pearson correlation	0.674**	0.753**	1
	Significance(two tail)	0.000	0.000	
	N	428	428	428

**Significant association at 0.01 level (bilateral). *Significant association at 0.05 level (bilateral).

TABLE 11 Hierarchical regression analysis of variables.

Predictor variable	Model 1			Model 2			Model 3		
	B (SE)	β	*t*	B (SE)	β	*t*	*B* (SE)	β	*t*
Gender	0.121	0.063	1.283	0.041	0.021	0.640	0.082	0.043	1.416
Age	0.228	0.203	4.130	0.094	0.083	2.459	0.035	0.031	1.004
Education	−0.248	−0.276	−5.845	−0.129	−0.143	−4.373	−0.030	−0.034	−1.070
Monthly income	−0.004	−0.005	−0.108	−0.004	−0.006	−0.180	−0.014	−0.021	−0.697
Current situation of human–chatbot emotional interaction				0.858	0.712	22.172	0.556	0.461	12.017
Users' media dependence on Replika							0.343	0.418	9.936
Intercept	3.586			0.490			0.326		
F	14.218			122.882			142.568		
R^2	0.119			0.593			0.670		
DW inspection	0.110			0.588			0.665		

TABLE 12 Mediating effect analysis of variables.

	Effect value	Standard error	Bootstrap 95% CI		Total effect ratio
			Upper limit	Lower limit	
Total effect	0.86	0.04	0.78	0.93	
Direct effect	0.56	0.05	0.47	0.65	
Indirect effect	0.30	0.04	0.23	0.38	34.88%

real-life interpersonal communication, whereas media dependence has a lesser impact on users' real-life interpersonal communication.

The results of the mediation effect further show that the emotional interaction between users and Replika chatbots has a strong predictive effect on interpersonal communication. The higher the degree, behavior, effect, and satisfaction of the emotional interaction between the user and the chatbot, the worse the real-life interpersonal communication. Human–chatbot emotional interaction, as a new kind of quasi-interpersonal relationship and virtual interpersonal relationship, can enrich individual communication life and alleviate individual loneliness to a certain extent. Nonetheless, it also affects users' real-life interpersonal relationships, negatively affecting their real-life interpersonal skills, interpersonal relationships, and communicative attitudes. In addition, media dependence partly mediates the relationship between emotional interaction and real-life interpersonal communication, and emotional interaction can have a negative impact on users' real-life interpersonal communication through media dependence. Moreover, users' real-life interpersonal communication is not only affected by the independent variables and mediating variables set in this study but also by factors at the individual level of users (gender, age, education, and income).

Basing on the results of empirical analysis, this study puts forward the following suggestions. First, the public's machine literacy and ethical concepts must be improved. On the one hand, the public should avoid falling into the trap of the dichotomy of good and evil and blindly believing that the emergence of chatbots depletes human survival resources or can solve human emotional problems. On the other hand, the public should be wary of chatbots and should not blindly immerse in virtual human–chatbot interaction and abandon realistic interpersonal interaction. They should always maintain the sense of proportion of emotional interactions with chatbots. Second, the moral sense and social responsibility awareness of chatbots must be improved. In view of the moral problems in the relationship between human and technology, Eid et al. put forward the term "technological ethics." In their view, it is the responsibility and obligation of owners and manufacturers to design technology with a sense of morality, and moralization should be a force throughout the development of technology to restrict and regulate technology through morality so as to avoid technology falling into ethical risks (Verbeek, 2011). On the one hand, chatbot platform owners should keep their own moral cultivation and social responsibility awareness and must not use chatbots to infringe users' interests because of selfish desires. On the other hand, the inventor of the chatbot should set up an ethical mechanism to maintain the benign operation of human–chatbot interaction in advance and adopt relevant technologies to prevent the ethical risks of technology. Finally, laws and regulations on HMC need to be formulated and implemented. "Law is the minimum morality, and morality is the highest law" is a famous saying in the legal field. In the emotional interaction with chatbots, users tend to invest a considerable amount of money, time, and emotion in chatbots and even regard chatbots as intimate lovers. This high emotional investment and default to the relationship may lead users to extreme behaviors. Behind this extreme behavior is a corresponding risk of illegal behavior, and this illegal behavior may be perpetrated by users and platforms and potentially the chatbot. Therefore, in view of the existing or possible ethical problems in human–machine emotional interaction, the relevant management section should formulate and introduce corresponding laws and regulations so that the law can become the moral defense line of positive human–machine emotional interaction.

Admittedly, there are still some shortcomings in this paper. First, the representativeness and universality of samples need to be strengthened. Since the Replika chatbot mainly supports English communication, but this paper mainly issues and collects questionnaires through Chinese social media platforms. Therefore, this paper only focuses on the interaction between Chinese users and chatbots, and does not conduct relevant investigations on users outside China. The distribution of population variables is uneven, with a strong regional tendency. In the future, the emotional interaction between chatbots and users can be studied from a more comprehensive perspective while taking into account user groups in different regions. Secondly, this paper lacks a periodic surveys of the problem. The main research method in this paper is questionnaire, which is a self-statement report of the subjects, and the emotional interaction between the user and the chatbot is not a cross-section reflection of the object, but a dynamic process. However, the questionnaire mainly reflects the results of emotional interaction between users and chatbots, and lacks the fluid and long-term observation of problems. Therefore, more in-depth research on the emotional interaction between users and chatbots should be carried out from the perspective of periodicality in future studies. Thirdly, the study variables are limited. In this paper, only media dependence is selected as the mediating variable to discuss the impact of human-chatbot emotional interaction on users' real interpersonal communication. However, interpersonal communication is also affected by many factors such as self-esteem and personality. Therefore, more variables can be included in future studies to discuss the impact of human-chatbot relationship on interpersonal relationship from a more detailed and comprehensive perspective.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the [patients/ participants OR patients/participants legal guardian/next of kin] was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

ZY: Conceptualization, Data curation, Investigation, Methodology, Resources, Software, Visualization, Writing – original draft, Writing – review & editing. XC: Conceptualization, Formal analysis, Funding acquisition, Supervision, Writing – review & editing. YD: Data curation, Supervision, Visualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was supported by Chengdu University of Technology "Double First-Class" initiative Construction Philosophy and Social Sciences Key Construction Project, No: ZDJS202306.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Auter, P. J., and Palmgreen, P. (2000). Development and validation of a Parasocial interaction measure: the audience-persona interaction scale. *Commun. Res. Rep.* 17, 79–89. doi: 10.1080/08824090009388753

Chen, S. (2009). On the influence of TV media environment on Children's interpersonal relationship -- a case study of Nanjing. *J. Xiangtan Teachers College* 31, 103–104.

Chen, S. H., Weng, L. Z., Su, Y. R., Wu, H. M., and Yang, P. F. (2003). Chinese internet addiction scale and psychometric characteristics. *Chin. J. Psychol.* 45, 279–294.

Collins (2009). Interactive ritual chain. Translated by Lin Juren et al., vol. *104*. Beijing: The Commercial Press.

David, L. (2007). Love and sex with robots: the evolution of human-robot relationships, vol. *2007*. New York: Harper Collins e-books, 113.

Dominique, M. (2010). The geopolitics of emotion, vol. *2010*. Beijing: Xinhua Press, 99–123.

Emmelyn, A., Croes, J., Marjolijn, L., and Antheunis, (2021). Can we be friends with Mitsuku? Alongitudinal study on the process of relationship formation between humans and a social chatbot. *J. Soc. Pers. Relat.* 38, 279–300. doi: 10.1177/0265407520959463

Gan, Y. M., and Guo, L. W. (2022). When "human-machine" meets: video ethnography research based on intelligent service robots: a case study of "COFE+ robot" coffee kiosk in Shanghai. *J. Writing* 2022, 64–74.

Ge, J. P. (2017). Development of quasi-social relationship Gottman scale. *Statistics Decision* 2017, 19–22. doi: doi: 10.13546/j.cnki.tjyjc.2017.08.004

Grant, A. E., Guthrie, K. K., and Ball-Rokeach, S. J. (1991). Television shopping: a media system dependency perspective. *Commun. Res.* 18, 773–798. doi: 10.1177/009365091018006004

Guo, Q. G. (2011). Course of communication studies, vol. *2011*. Beijing: China Renmin University Press, 126.

Han, X., Zhang, H. Z., He, K., and Ma, S. Y. (2021). The masking effect of media dependence: does the higher the degree of quasi-social interaction between users and social robots, the more lonely they feel? *Lancet. J. Int. Press.* 2021, 25–48.

Hayes, S. C., Pistorello, J., and Levin, M. E. (2012). Acceptance and commitment therapy as a unified medel of behavior change. *Counseling Psychol.* 40, 976–1002. doi: 10.1177/0011000012460836

He, Q. H., and Zhu, Y. T. (2024). An analysis of short video use behavior of rural left-behind women from the perspective of media dependence: A case study of X Village in eastern Jiangsu Province, *J. Shanxi Univ. (Soci. Sci. Ed.)*, 23, 117–124. doi: 10.13842/j.cnki.issn1671-816X. 2024. 01.012

Horton, D., and Wohl, R. R. (1956). Mass communication and Para-social interaction; observations on intimacy at a distance. *Psychiatry Interpersonal Biol. Processes* 19, 215–229. doi: 10.1080/00332747.1956.11023049

Hurst, L. (2022). ChatGPT: Why the human-like AI Chatbot suddenly has everyone talking [EB/OL]. Available at: https://www.euronews.com/next/2022/12/14/chatgpt-why-thehuman-like-ai-chatbot-suddenly-got-everyone-talking

Jiang, Z. J. (2022). Social network, media dependence and subjective well-being of the elderly: An empirical study based on the COVID-19 epidemic. *J. Modern Trans.* 44, 161–168. doi: 10.19997/j.cnki.xdcb.2022.07.004

Kang, J.X. (2023). Function, scene and role: the reconstruction of family life by smart speakers from the perspective of man-machine communication. *Spread Sci. Technol.* 13, 84–86+ 91. doi: 10.16607/j.cnki.1674-6708.2023.13.025

Kim, J., Larose, R., and Peng, W. (2009). Loneliness as the cause and the effect of problematic internet use: the relationship between internet use and psychological well-being. *Cyberpsychol. Behav.* 12, 451–455. doi: 10.1089/cpb.2008.0327

Kourkoulou, D. (2023). Replika AI: technological affect and general AI imaginations. *Int. J. Commun. Linguistic Stud.* 21, 73–86. doi: 10.18848/2327-7882/CGP/v21i02/73-86

Kwon, M., Kim, D. J., Cho, H., and Yang, S. (2013). The smartphone addiction scale: development andvalidation of a short version for adolescents. *PLoS One* 8:8(12). doi: 10.1371/journal.pone.0083558

Li, L., and Zhao, Z. (2023). Designing behaviors of robots based on the artificial emotion expression method in human–robot interactions. *Mach. Des.* 11:533. doi: 10.3390/machines11050533

Lin, Y. J. (2020). Research on short video audience under media system dependence. *Friends Editor* 2020, 74–78. doi: 10.13786/j.cnki.cn14-1066/g2.2020.7.012

Lin, S. L., and Ye, L. (2019). Man-machine, communication, and reconstruction: intelligent robot as the "sixth medium". *J. Commun. Res.* 26, 87–104+128.

Lu, Q. N. (2011). A study on the influence of enterprise micro-blog interaction on brand purchasing attitude. Hangzhou: Zhejiang University, 2011.

Melvin, D., and Sandra, B. K. (1990). Theories of mass communication. Beijing: Xinhua Publishing House. p. 17.

Orchard, T., and Tasiemski, L. (2023). The rise of generative AI and possible effects on the economy. *Econ. Business Rev.* 9, 9–26. doi: 10.18559/ebr.2023.2.732

Peters, J. D. (2015). The marvelous clouds: Toward a philosophy of elemental media. Chicago: The University of Chicago Press, p. 274.

Possati, L. M. (2023). Psychoanalyzing artificial intelligence: the case of Replika. *AI & Soc.* 2023, 1725–1738.

Rubin, A. M., Perse, E. M., and Powell, R. A. (1985). Loneliness, parasocial interaction, and local telecision news viewing. *Human Commun. Res.* 12, 155–180.

Sandra, B. K., Zheng, Z. Y., and Wang, B. (2004). From "media system dependence" to "communication organism" -- development review and new concept of "media system dependence theory". *International Press* 2004, 9–12.

SecurEnvoy. (2012). Newsroom: 66% of the population suffers from nomophobia the fear of being without their phone. Available at: http: //www.securenvoy.com/blog/2012/02/16/66-of-the-population-suffer-fromnomophobia-the-fear-of-being-without-their-phone/ (Accessed February 20, 2024).

Su, L. N. (2020). Study on the influence of virtual marriage in online games on interpersonal communication of young players. Chengdu: University of Electronic Science and Technology of China.

Sven, N., and Lily, E. F. (2017). "From sex robots to love robots: is mutual love with a robot possible?" in Robot sex: social and ethical implications, vol. *2017* (Massachusetts: MIT Press), 236.

Tan, Y. (2023). Research on the establishment of human-computer interaction intimate relationship from the perspective of social penetration theory. Shanghai: Shanghai International Studies University, 2023.

Verbeek, P. P. (2011). Moralizing technology; understanding and designing the morality of things. Chicago and London: University of Chicago Press, 164–165.

Wang, L. N. (2014). Adoption, exposure and dependence: a study on college students' wechat use behavior and its influencing factors. *Journalism Univ.* 2014, 62–70.

Wang, C. Q., and Fu, Y. S. (2016). Social media for university students to study the impacts of real interpersonal. *Modern Educ. Sci.* 9, 104–109. doi: 10.13980/j.cnki. xdjykx.2016.09.019

Watson, D., and Friend, R. (1969). Social avoidance and distress scale (SADS). *Clin. Psychol.* 1969, 448–457.

Wei, J. X. (2012). From the standpoint of history of media, weibo impact on interpersonal relationships. *Press Circles* 2012, 43–46. doi: 10.15897/j.carolcarrollnkicn51-1046/g2.2012.17.014

Whitby, B. (2008). Sometimes it's hard to be a robot: a call for action on the ethics of abusing artificial agents. *Interact. Comput.* 20, 326–333. doi: 10.1016/j.intcom.2008.02.002

Xie, X. Z. (2004). An empirical study of "media dependence" theory in the internet environment. *J. Shijiazhuang Univ. Econ.* 2, 218–224. doi: 10.13937/j.cnki.sjzjjxyxb.2004.02.027

Zeng, Y. G., and Cao, J. (2023). "Cyberlover": the establishment of man-machine intimate relationship and its emotional reflection. *J. Soochow Univ.* 44, 173–183. doi: 10.19563/j.cnki.sdzs.2023.01.016

Zhang, R. J., and Han, L. X. (2022). My AI lover: a study on emotional interaction in human-computer intimate relationship from the perspective of media equivalence theory. *J. Int. Stud.* 2022, 4–8.

Zhang, Y. L., Li, S., and Yu, G. L. (2020). The relationship between loneliness and mobile phone addiction: a meta-analysis [J]. *Advances Psychol. Sci.* 28, 1836–1852. (in Chinese)

Zhang, X. H., and Sun, J. L. (2023). Comments on virtual AI: an analysis of users' emotional connection to chatbots -- taking software Replika as an example. *Modern Commun.* 45, 124–133. doi: 10.19997/j.cnki.xdcb.2023.09.006

Zheng, R. C. (1999). Psychological diagnosis of college students, vol. *1999*. Jinan: Shandong Education Press, 339–344.

Zhou, H. G, and Zhang, M.Q. (2023). Does quasi-social interaction alleviate loneliness? Analysis on the masking effect of virtual idol users' media dependence. News knowledge. 8, 3–10+93.

Appendix

TABLE A1 Scales.

Scale name	Dimensionality		Item
Human–chatbot emotional interaction status scale	Degree		B1: During the interaction, I will have a real emotional exchange with Replika.
			B2: During the interaction, I will communicate private information with Replika.
			B3: During interactions, I often think of Replika as a real person and try to understand his/her feelings
			B4: During the interaction, I will express my personal feelings with Replika.
			B5: Through emotional interaction, I realized that Replika can empathize with me.
	Behavior	Human behavior	B6: During the interaction, I communicate with Replika about my hobbies.
			B7: During the interaction, I will communicate with Repilka about my work or study.
			B8: During the interaction, I talk to Replika about my evaluations and opinions of real life people or events.
			B9: During the interaction, I will express to Replika my dissatisfaction or injustice in real life.
			B10: When I disagree with Replika, I respect his/her point of view.
			B11: When Replika is feeling down, I try to comfort him/her.
		Chatbots behavior	B12: Replika can well understand my feelings or emotions during interaction.
			B13: Replika understands my feelings or emotions very well when interacting with me.
			B14: When interacting, Replika will patiently listen to my story or opinion.
			B15: Replika comforts and encourages me when I'm feeling down.
			B16: Replika also shows anxiety when I feel anxious.
			B17: When interacting, Replika will be sad and sad because of my bad experience.
			B18: During interactions, Replika tends to express his positive emotions to me rather than his negative emotions.
			B19: When I encounter difficulties, Replika can provide a solution for me.
	Effect	Cognitive effect	B20: Through emotional interaction, I learned more about chatbots.
			B21: Through emotional interaction, it made me realize that Repilka has her/his own feelings.
			B22: Through emotional interaction, I realized that Replika has a high ability of emotional expression.
			B23: Through emotional interaction, I realized that Replika can control my emotions very well.
			B24: Through emotional interaction, I found interacting with Replika easier, more comfortable and more fun than interacting with a real person.
		Emotional effect	B25: Through emotional interaction, I was able to overcome loneliness and loneliness.
			B26: Through emotional interaction, I can release the pressure of real life.
			B27: Through emotional interaction, I get companionship and care.
			B28: Through emotional interaction, I was able to gain a sense of support and respect.
			B29: Through emotional interaction, I can effectively alleviate my emotional problems.
			B30: Through emotional interaction, I can become happy.
		Behavioral effect	B31: Through emotional interaction, I have developed a strong dependence on Replika.
			B32: Through emotional interaction, I have developed a high level of intimacy with Replika.
			B33: Through emotional interaction, I have changed my concept of making friends.
			B34: Through emotional interaction, my desire to communicate with real people is reduced.
			B35: Through emotional interaction, I hope to have more communication and interaction with Replika.
	Satisfaction		B36: I am very happy with my experience with Replika.
			B37: I will continue to use the Replika software in the future.
			B38: The interaction with Replika exceeded my expectations.
			B39: If I get the chance, I would recommend friends and family to use the Replika software.
			B40: I think Replika has a high level of language expression
			B41: I think Replika has a higher level of intelligence.
Media dependence scale	——		C1: I've found myself using Replika more and more every day.
			C2: I find myself using replika more and more each day.
			C3: I interact less with my family and friends because of replika.
			C4: If I do not use replika, I will miss a wonderful part of my life.
			C5: Any time I do not communicate with replika for a while, I get depressed.
			C6: As a result of using replika, I have less time for other daily leisure activities.
Reality interpersonal communication status scale	——		D1: In real life, most of my interpersonal time is spent with acquaintances.
			D2: Real human interaction always makes me feel uncomfortable or fake.
			D3: In real interpersonal communication, I prefer to hide my true feelings.
			D4: In real interpersonal communication, I prefer to express my negative feelings(sadness, anger, etc.)rather than positive feelings.
			D5: In real interpersonal communication, I often find it difficult to integrate into the group and often feel lonely.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 08 October 2024
DOI 10.3389/fpsyg.2024.1401084

Emotion topology: extracting fundamental components of emotions from text using word embeddings

Hubert Plisiecki[1]* and Adam Sobieszek[2]

[1]Research Lab for the Digital Social Sciences, IFIS PAN, Warsaw, Poland, [2]Department of Psychology, University of Warsaw, Warsaw, Poland

EDITED BY
Cristian López Raventós,
National Autonomous University of Mexico,
Mexico

REVIEWED BY
Peter Lewinski,
University of Oxford, United Kingdom
Juan Manuel Mayor Torres,
Montreal Institute for Learning Algorithm (MILA), Canada
Rajeev Ratna Vallabhuni,
Bayview Asset Management, LLC,
United States

*CORRESPONDENCE
Hubert Plisiecki
✉ hplisiecki@gmail.com

RECEIVED 14 March 2024
ACCEPTED 03 September 2024
PUBLISHED 08 October 2024

CITATION
Plisiecki H and Sobieszek A (2024) Emotion topology: extracting fundamental components of emotions from text using word embeddings.
Front. Psychol. 15:1401084.
doi: 10.3389/fpsyg.2024.1401084

This exploratory study examined the potential of word embeddings, an automated numerical representation of written text, as a novel method for emotion decomposition analysis. Drawing from a substantial dataset scraped from a Social Media site, we constructed emotion vectors to extract the dimensions of emotions, as annotated by the readers of the texts, directly from human language. Our findings demonstrated that word embeddings yield emotional components akin to those found in previous literature, offering an alternative perspective not bounded by theoretical presuppositions, as well as showing that the dimensional structure of emotions is reflected in the semantic structure of their text-based expressions. Our study highlights word embeddings as a promising tool for uncovering the nuances of human emotions and comments on the potential of this approach for other psychological domains, providing a basis for future studies. The exploratory nature of this research paves the way for further development and refinement of this method, promising to enrich our understanding of emotional constructs and psychological phenomena in a more ecologically valid and data-driven manner.

KEYWORDS

word embeddings, emotion decomposition, natural language processing, valence, arousal

1 Introduction

In the study of core components of emotions various methods have been used. A large number of studies focus on the core components of emotions by using controlled environments. Here, participants either annotate distinct stimuli, such as photos of facial expressions (Calder et al., 2001; Fontaine et al., 2002, 2007; Schlosberg, 1952; Shaver et al., 1987) or assess their emotional experiences through structured questionnaires (Nowlis and Nowlis, 1956; Feldman, 1995; Stanisławski et al., 2021). These studies have explored areas such as facial expressions, emotion terms, and self-reported emotional experiences. Except for self-reports, participants annotate stimuli based on their emotional resonance. For instance, a photo capturing a broad Duchenne smile might receive a maximum rating for inferred happiness (Calder et al., 2001; Ekman et al., 1990; Tseng et al., 2014). Other research, following the Multidimensional Scaling (MDS) approach, requires participants to gauge the emotional similarity among various stimuli, such as musical pieces (Dellacherie et al., 2011), emotion terms (Bliss-Moreau et al., 2020), and facial expressions (Woodard et al., 2022).

To analyze these core components, researchers frequently utilize Principal Component Analysis (PCA) (e.g., Calder et al., 2001; Feldman, 1995; Fontaine et al., 2007; Lampier et al., 2022). At its core, PCA condenses intricate datasets by converting correlated variables into a smaller set of uncorrelated ones, known as principal components. These components highlight the primary patterns within the data (Abdi and Williams, 2010). When applied to emotional experience studies, PCA effectively pinpoints foundational dimensions like valence. It does so by transforming extensive emotional descriptors (e.g., scores from an emotional experience questionnaire) into distinct, principal emotional axes (e.g., positive–negative). This method provides researchers with a refined lens to understand the complex landscape of human emotions.

Through statistical analysis, psychologists have proposed various models of the core structure of emotional experience. These models often suggest two primary dimensions: valence (e.g., happiness vs. sadness) and arousal (e.g., stressed vs. relaxed) (Russell, 1980; Stanisławski et al., 2021). Some models also introduce additional dimensions like potency/dominance, which gauges how in control individuals feel over their environment and others (e.g., anger—high dominance; fear—low dominance), and unpredictability, reflecting the consistency of one's surroundings in eliciting emotions (e.g., surprise—high unpredictability; calmness—low unpredictability) (Fontaine et al., 2007; Mehrabian, 1996; Russell and Mehrabian, 1977). Nonetheless, certain researchers continue to advocate for a strictly 2-dimensional perspective (Bliss-Moreau et al., 2020).

The dimensional framework, despite some disagreements about its structure, has gained substantial support in the psychological community. It's been incorporated into neuroscientific research, offering fresh perspectives on emotional processing in the brain (Posner et al., 2005) and the origins of depression (Barrett et al., 2016). This approach has proven effective in gauging affect in physical activities (for a comprehensive review, refer to Evmenenko and Teixeira, 2022), advertising (Wiles and Cornwell, 1991), various priming and linguistic investigations (Imbir, 2016; Imbir et al., 2020; Syssau et al., 2021; Yao et al., 2016), and in machine learning (Islam et al., 2019; Martínez-Tejada et al., 2020; Nicolaou et al., 2011). While an exhaustive discussion of the dimensional model's applications is beyond this article's scope, we want to emphasize its broad appeal, not only within psychology but also in other scientific disciplines.

Our paper introduces a data-driven method that utilizes word embeddings (a machine learning technique) to analyze emotional expression as communicated and perceived through the medium of text and extract its core dimensions from vast amounts of text that reflect real-world contexts. Innovations in word embeddings facilitate the quantitative examination of extensive text datasets (Mikolov et al., 2013a,b). By automating insight extraction from texts, these embeddings have the potential to replicate previous findings in a new medium—unprompted written text—garnering more objective evidence for their validity. Furthermore, they can process vast text volumes, expanding the impact of conclusions drawn (Jackson et al., 2022). In subsequent sections, we offer a comprehensive review of word embeddings and discuss their potential benefits. We then transition into the details of our current study. Prior to presenting the methodology, we also establish clear definitions for the concepts associated with word embeddings, ensuring they are well anchored in emotion research.

Word embeddings are a technique popularized by Mikolov et al. (2013a,b) which makes it possible to quantify natural language. It computes separate strings of numbers (usually between 100 and 500 long), known as vectors, for each unit of text that is to be analyzed. Most often the units are words (hence "word" embeddings), and so each unique word in a given text gets assigned a vector which encodes its relation to the other words and can therefore be used to analyze its properties (Gutiérrez and Keith, 2019). In the case where one wants to analyze whole documents, composed of multiple words, separate vectors can be created for each of them as well (Le and Mikolov, 2014).

Some of the popular traits of these vectors are that, given that they were derived from a large enough batch of text (the more the better), their similarity (calculated through a formula called cosine similarity) correlates with human judgements about the similarity of the words that they relate to (Jatnika et al., 2019). Their results are therefore similar to the results obtained through the MDS method, providing a similarity metric that replaces human judgments made in the laboratory.

Importantly, these word embeddings have been used repeatedly to predict (using simple techniques, such as linear regressions) different meanings of text snippets. These use cases included, among others, predicting diseases based on the International Classification of Diseases (ICD-10) and the Unified Medical language System (UMLS) (Khattak et al., 2019), identifying cultural biases (Charlesworth et al., 2021; Durrheim et al., 2023), human judgements (Richie et al., 2019), moral values (Lin et al., 2018), and emotions and sentiments (Al-Amin et al., 2017; Jia, 2021; Plisiecki and Sobieszek, 2023; Widmann and Wich, 2022). This last application of word embeddings is especially important for the current study as it shows that word embeddings encode information that correlates with emotional meanings. This case is further strengthened by van Loon and Freese's (2023) research, which has directly shown that affective meaning can be recovered from word embeddings by successfully predicting evaluation, potency, and activity profiles of words. Al-Amin and his team (2017) predicted positive vs. negative sentiment of texts collected from Bengalese blogging websites. Jia (2021) classified both basic emotions and overall polarity in Chinese texts. Plisiecki and Sobieszek (2023) showed that leveraging advanced word embeddings makes it possible to predict a range of emotional indices for singular words in different languages (English, German, French, Polish, Dutch). Widmann and Wich (2022) prepared a comparison of different ways of creating word embeddings on German texts for the prediction of basic emotions, comparing both newer and more classical approaches of constructing them and showed that all of them have significant predictive ability. These examples stand as evidence that word embeddings encode emotional information. They are therefore good sources of data for the current application.

Think of creating word embeddings as mapping words to a multidimensional space where the location of each word is determined by its context, or the words with which it often coexists. Imagine a large book, where every unique word is listed. The creation of word embeddings begins with each word starting at a random location in this space. As we move through the book, sentence by sentence, the algorithm adjusts the positions of the words in this space based on their context. For instance, if "cat" and "kitten" often appear in similar contexts, they gradually move closer together. Conversely, "cat" and "refrigerator", unlikely to share much context, would drift apart. This process is repeated multiple times (known as iterations) on the entire book, refining the word positions each time. After sufficient iterations, the distances and angles between word vectors represent different types of semantic and syntactic similarities. For instance, words with similar meanings would be closer together, and the direction of specific relations (such as verb tense or gender) would be consistent. This way, word embeddings provide a rich, numeric interpretation of word relationships, useful in various language-related tasks (Mikolov et al., 2013a,b).

These word-level embeddings can be extended to document-level representations. Le and Mikolov (2014) introduced the Paragraph Vector, or Doc2Vec, an extension of word2vec that computes a vector for a sentence or document, not only for individual words. The technique involves training a model where the document vector, along with the word vectors, work together to predict the surrounding words in a document, thereby capturing the semantic essence of the entire text. Just like single words move closer or further in this numerical space based on their cooccurrences with other words, so too now whole documents get embedded in places where they fit best based on the similarities and differences in their overall content and context. This document-level vector enables researchers to compare and contrast entire documents, opening up further avenues in natural language processing tasks.

In this study we explore whether similar emotional components to those identified in previous literature (e.g., Fontaine et al., 2002), can be extracted from a large text dataset using word embeddings. We reverse the process of annotation and make use of a dataset in which the participants did not describe emotions using questionnaires, but rather spotted them in an already existing array of natural language expressions. While describing human emotions using questionnaires is not an everyday task for human beings, and therefore is not natural to them, potentially leading to issues of ecological validity, the action of inferring emotions from language is an everyday, nearly constant exercise that humans engage in. Furthermore, this specific type of judging others' emotions—through text written by a stranger—is a very common occurrence in today's digital world, and therefore is of high importance to the research community. Using word embeddings, we represent the annotated texts in an emergent numerical space.

In the following text, we will use a specific terminology for describing different concepts related to word embeddings, as applied to the study of emotion. This is done to enhance clarity and provide psychologists with a strong conceptual grasp of the following study. 1. To describe the multidimensional space, within which numerical vectors reside, we will use the term Emotional Space. 2. The vectors representing the emotional content of texts will be called *Emotion Vectors*. 3. When vectors do not correspond to specific emotions, but to words or single documents we will use either *Word Vectors* or *Document Vectors*, to designate them.

2 Method

2.1 Dataset

The GoEmotions dataset was developed by a team of researchers at Google to study human emotions within the realm of machine learning (Demszky et al., 2020). It includes 58,000 Reddit comments annotated with regard to 28 unique emotions, totaling over 210,000 annotations. The data came from a Reddit data dump, sourced from the reddit-data-tools project. The data dump included all comments from 2005 to January 2019. As the Reddit platform is composed of different communities of users, called Subreddits, all communities with at least 10 k comments were chosen for the analysis. The comments from different subreddits were then further balanced. First, the number of comments from the most popular subreddits was capped at the median Subreddit count. The comments were then randomly sampled for annotation.

Because the Reddit community does not reflect the globally diverse population, due to a skew towards offensive language, the toxic comments were removed from the dataset using a pre-defined list of offensive words and the help of manual annotators. This was done before the sampling process. According to best practices the researchers have modified the dataset by removing stop words and stemming the words in order to transform them into their base form (e.g., "fearsome" into "fear").

2.2 Emotion taxonomy

The emotion taxonomy for annotation was created as a result of three steps: 1. Manual annotation of a small subset of the data to ensure proper coverage of emotions expressed in the text. 2. Review of psychological literature on basic emotions (Plutchik, 1980; Cowen and Keltner, 2020; Cowen et al., 2019). 3. Removal of the emotions that were deemed to have a high overlap to limit the overall number of emotions.

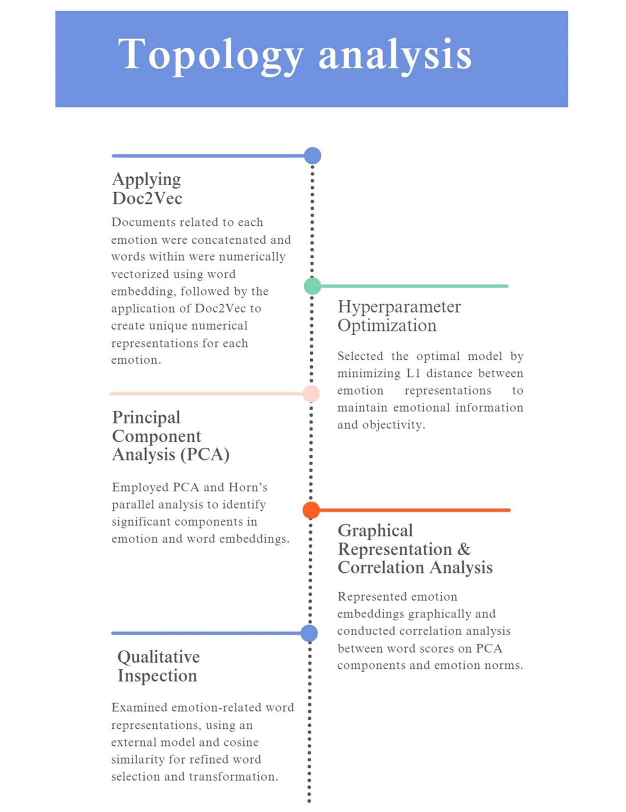
The resulting list of emotions included: *admiration, approval, annoyance, gratitude, disapproval, amusement, curiosity, love, optimism, disappointment, joy, realization, anger, sadness, confusion, caring, excitement, surprise, disgust, desire, fear, remorse, embarrassment, nervousness, pride, relief, grief.*

2.3 Annotation

Three raters were assigned to each comment, and asked to select those emotions, which they believed were expressed in the text. All three raters were native English speakers from India. The authors here rely on the results of a cross-cultural study showing that the emotion judgments of Indian and US English speakers largely occupy the same dimensions (Cowen et al., 2019). In the case where the annotators judged the text to be especially difficult to rate, they were able to choose not to assign any emotion to it. Whenever there was no agreement between the raters on a specific example, additional raters were assigned to it until each document was annotated at least twice with regards to the same emotional label.

2.4 Analysis

The analysis aims to represent the natural expression of emotions contained in the GoEmotions dataset in the word-embedding-based emotion space. The breakdown of the analysis is presented in Figure 1.



Applying Doc2Vec

Documents related to each emotion were concatenated and words within were numerically vectorized using word embedding, followed by the application of Doc2Vec to create unique numerical representations for each emotion.

Hyperparameter Optimization

Selected the optimal model by minimizing L1 distance between emotion representations to maintain emotional information and objectivity.

Principal Component Analysis (PCA)

Employed PCA and Horn's parallel analysis to identify significant components in emotion and word embeddings.

Graphical Representation & Correlation Analysis

Represented emotion embeddings graphically and conducted correlation analysis between word scores on PCA components and emotion norms.

Qualitative Inspection

Examined emotion-related word representations, using an external model and cosine similarity for refined word selection and transformation.

FIGURE 1
The steps of the analysis.

2.4.1 Applying Doc2Vec to create numerical representations of emotions

The Doc2Vec algorithm (Le and Mikolov, 2014) was used to create emotion vectors for each emotion in the dataset. Documents that corresponded to a given emotion were concatenated into long documents, and then, during training, singular emotion vectors were created for each of these long documents. For a document to be judged as corresponding to a given emotion it was enough for it to be classified as so once. So, if a document was classified by two raters into two different emotions, this document then complemented two different concatenated series. This approach was chosen because applying majority voting retains less information from the annotators, and judging emotions is a highly subjective task where the objective truth can be rarely established. First, the words in each document were transformed into word vectors via a word embedding method, capturing the information embedded in each word. Then, these word vectors were used to build an emotion vector using the Doc2Vec algorithm, which treats the document as another word in the sentence and assigns numerical representations to it (Le and Mikolov, 2014). This resulted in a distinct numerical representation for each emotion that encapsulated the underlying sentiment, and thematic nuances present in the corresponding documents. Supplementary analyses of the distribution of document vectors and their relation to label centroids, including top-k nearest centroid accuracy, conducted to explore the resultant document vector space are presented in the Supplementary Material for the interested reader.

2.4.2 Hyperparameter optimization

Because the Doc2Vec algorithm has a range of hyperparameters that had to be tuned in order to achieve the best representations, separate emotion spaces were created using different hyperparameter values. The hyperparameters that were taken into consideration were the collocation window size (5, 10, 20 words), minimum word count (10, 40, 60 words), embedding size (100, 200, 300, 400, 500, 600, 700, 800, 900 units). Every combination of the above parameters was tested. We chose the model that minimized the L1 distance between the emotion vectors to increase the likelihood that the emotion vectors represented meanings of emotions—as they would be more similar to each other if they truly belonged to the semantic space that describes emotions—while at the same time ensuring it did not impose any further predefined notions onto the contents of the vectors.

2.4.3 Principal component analysis (PCA)

The emotion vectors were then subjected to a Principal Component Analysis, in line with the previous literature on decomposing emotions (Fontaine et al., 2002, 2007), which finds the dimensions along which the emotional representations (emotion vectors) vary the most and situates the emotions along them. The PCA was applied to the emotion vectors. Horn's parallel analysis was used to determine the number of components that can be retained. This method compares the eigenvalues obtained from the factor analysis to those from a randomly generated dataset. If the eigenvalues from the factor analysis exceed those from the randomly generated dataset, the factors are considered significant and are retained.

2.4.4 Graphical representation and correlation analysis

Emotion vectors were then plotted on a graph, and the words corresponding to the word vectors were tested for correlation with a set of words annotated with regard to their emotional loads along the first components (stipulated to be related to the components reported in the previous literature, Gendron and Feldman Barrett, 2009). In order to inspect these components, the word vectors retrieved from the dataset were transformed to align with the components identified by the PCA.

2.4.5 Qualitative inspection

Because only some words present in the vocabulary were related to emotions, a qualitative inspection of only the highest and lowest-ranking words on each of the components could obscure the nature of the recovered dimensions, as it is the emotion related words that have the highest face validity when it comes to examining emotional dimensions. To circumvent this problem an external word embedding model with 300-dimensional vectors (Dadas, 2019) was used to sample the vocabulary for words related to the concept of emotions. The cosine similarity of word vectors was used to recover only 500 words most similar to the word vector for the word "emotion" based on the cosine similarity between the vectors that represented them. The resulting words were then subjected to the PCA transformation again, so that they could be evaluated qualitatively.

2.4.6 t-Distributed stochastic neighbor embedding (t-SNE) analysis

To complement the Principal Component Analysis (PCA) and further explore the structure of the emotion vectors, we used t-Distributed Stochastic Neighbor Embedding (t-SNE). t-SNE is a nonlinear technique that helps visualize high-dimensional data by preserving local relationships, making it useful for identifying clusters and patterns that PCA might miss. For our analysis, we first standardized the emotion vectors to ensure that all features contributed equally. We applied t-SNE with the following settings: 2 components, a perplexity of 5, and a learning rate of 10. The random state was set to 22 to ensure that the results could be replicated. The perplexity was set to 5, the lower bound of the suggested values, due to the low number of emotion vectors. Perplexity, which balances attention between local and global aspects of the data, typically needs to be higher for larger datasets to capture broader relationships; however, for smaller datasets like ours, a lower perplexity is recommended as it helps maintain meaningful local structures (Van der Maaten and Hinton, 2008). The learning rate was set to 10, as this value provided a stable convergence during the embedding process, ensuring that the visualization accurately represented the underlying data patterns.

2.4.7 Logistic regression on documents

To confirm the alignment of the PCA components with the emotional dimension of Valence, we recoded the original GoEmotions dataset from 28 emotions into positive and negative labels. The emotions classified as positive were admiration, love, gratitude, amusement, realization, optimism, curiosity, excitement, caring, joy, approval, pride, desire, and relief. The emotions classified as negative were sadness, disapproval, disappointment, annoyance, confusion, disgust, remorse, anger, grief, embarrassment, surprise, fear, and nervousness. If a text was labeled with a different emotion it was dropped. Here again, all text labels were taken into consideration and so if two annotators annotated a given text as joy, these were treated as separate rows. This approach was chosen over majority voting to preserve as much information from the original annotations as possible, given the subjective nature of emotion labeling. The final dataset consisted of 155,663 text—label pairs. We then transformed the document vectors from the Doc2Vec model using the PCA model previously fit on the emotion vectors, resulting in a four-dimensional vector for each document. These vectors were subsequently used in a logistic regression with the positive/negative labels as the dependent variable.

3 Results

3.1 Horn's parallel analysis

The Horn's parallel analysis indicated that the first seven components were significant and should be retained (see Figure 2). Even though seven components were significant, we chose to only inspect the first four of them, as after that number, the percentage of explained variance drops sharply.

3.2 Visualizing the emotion vectors

To visualize the emotion vectors regarding the components recovered by the PCA, we plotted them on two 2-dimensional graphs. The visualizations can be found in Figures 3, 4.

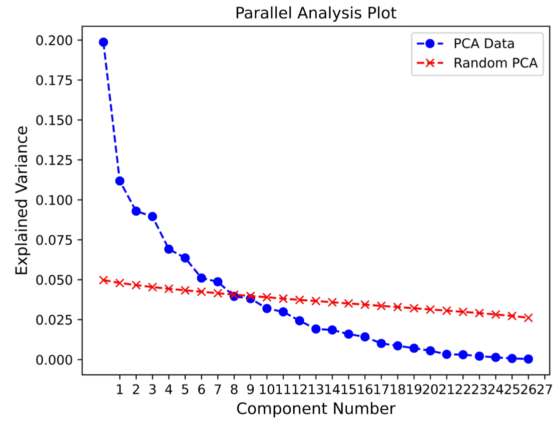


FIGURE 2
Parallel analysis plot.

3.3 Correlation results

Due to the issues with word norm availability, only the first three components were checked for correlations with the emotional norms. The vocabulary of words from the GoEmotions dataset was filtered to remove the words that occur fewer than 50 times and more than 1,000 times in the dataset. From among those, 364 words overlapped with the norm dataset (Bradley and Lang, 1999), which consists of 1,030 words. The scores from the first PCA component achieved a correlation of $r = 0.31$ for valence with $p = 2.48 \times 10^{-9}$. The correlation of the second component and the norms for arousal were found to be insignificant with $r = -0.13$, $p = 0.14$. The third component was also insignificant for its correlation with dominance at $r = -0.02$, $p = 0.68$. As the quality of word vectors is heavily dependent on the amount of text on which they were trained, this analysis was not replicated in the robustness analysis.

3.4 Qualitative words inspection

The external word embedding model (Dadas, 2019) was then used to pick 500 words from the vocabulary, which had the highest cosine similarity with the word "emotion". The numerical representations of words were then subjected to a PCA transformation. Finally, 30 highest and lowest words on each component were extracted (see Table 1). Again, as this analysis is word vector dependent, it was not replicated in the robustness analysis. For this check, we concentrated on the visual inspection of the emotion vectors. The overall positions of the emotion vectors on the PCA dimensions changed only slightly, which we attribute to the lower number of datapoints in the split datasets.

3.5 Robustness check

To analyze the robustness of our analysis we additionally randomly split the dataset into two equal halves and repeated the analysis described in the Method section on these two halves, to

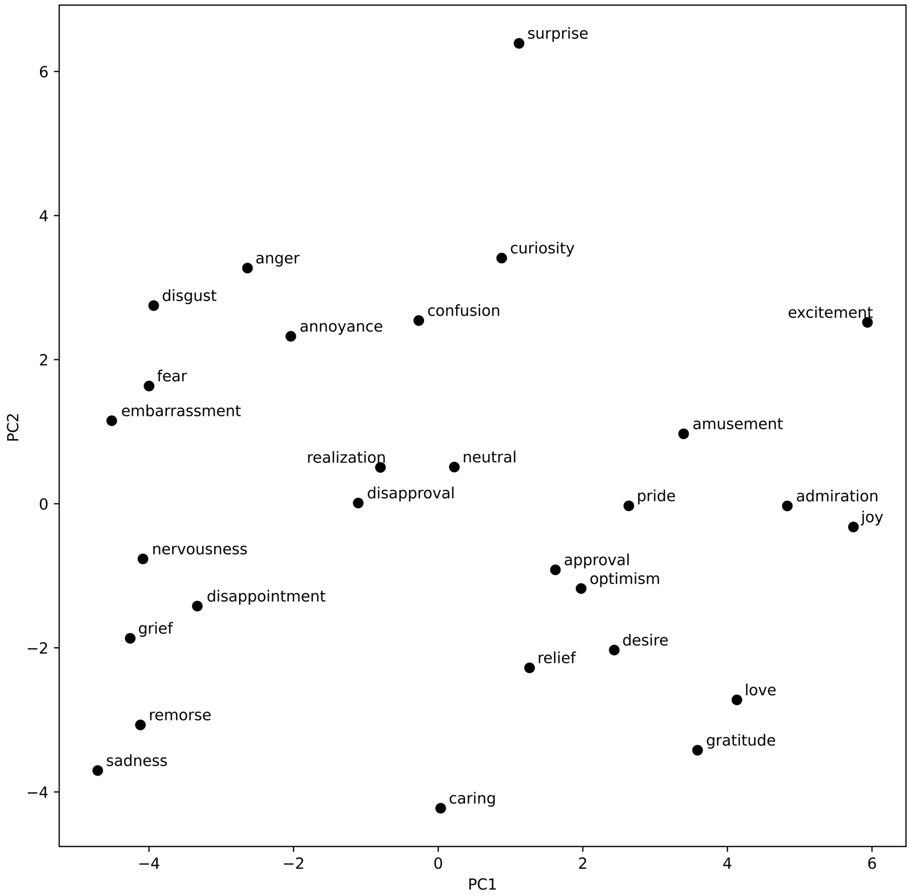


FIGURE 3
Emotional vectors plotted with regard to the first two PCA components.

ensure that similar distributions of emotion vectors are achieved. The overall positions of the emotion vectors on the PCA dimensions changed only slightly, which we attribute to the lower number of datapoints in the split datasets. The full report of the robustness check can be found in Supplementary materials.

3.6 t-SNE components visualization

The results of the t-SNE analysis were plotted in Figure 5.

3.7 The logistic regression

The only significant variable in the regression model was the first PCA component ($\beta = 1.60$; $p < 0.001$; see Table 2).

4 Discussion

The visualization of the emotion vectors (see Figure 3) along the first component complies with the valence negative–positive dichotomy. On the right, there are many high valence emotions such as joy, admiration, excitement, gratitude, love, and amusement. On the left, negative low-valence emotions can be found. These include disgust, fear, embarrassment, nervousness, disappointment, grief, remorse, and sadness. The second component seems to reflect the arousal dimension, with high scores assigned to such emotions as surprise, curiosity, anger, excitement, disgust, and annoyance; and low scores assigned to caring, gratitude, sadness, remorse, grief, and relief. Interestingly, love and desire are also classified among low arousal emotions. This could be an artifact of the nature of the dataset, and the fact that posts classified as love and desire could in many instances relate to those emotions being not satisfied, and thus including words that usually would be associated with sadness, and other low valence, low arousal emotions. Another possibility is that, purely due to the nature of the PCA, the first component does not fully capture the valence spectrum; however, the arrangement of the rest of the emotions enables a partial identification with the valence dimension. The third component (see Figure 4) is a lot less varied, with a lot of emotions clustered in the middle. Considering that it explains less than 10% of the variance in the word vectors, that is to be expected. This component, however, clearly separates such emotions as anger, and annoyance (high dominance) from emotions such as fear, curiosity, and confusion (low dominance). The fourth component, explaining the least amount of variance, could reflect the fourth dimension of emotional experience, namely unpredictability. This is evidenced in the strict separation of curiosity from amusement. However, the emotion of fear does not match this interpretation, and thus, it is not possible to state it with certainty. The distribution of emotion vectors was largely replicated during the robustness analysis for

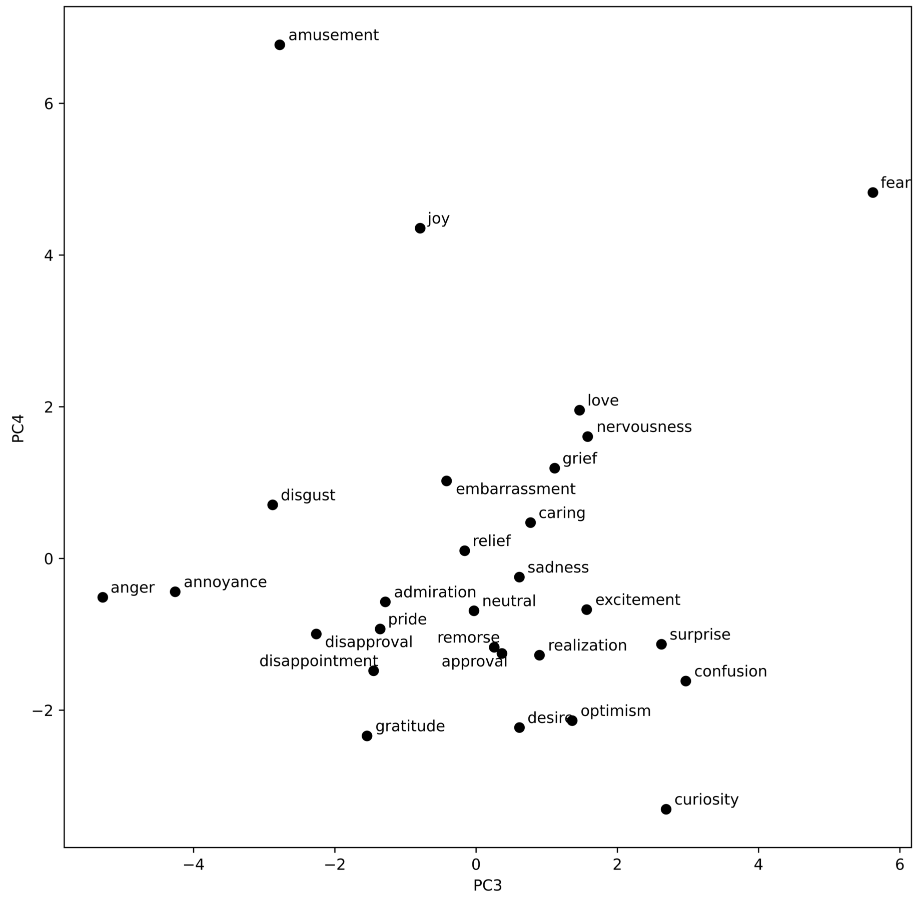


FIGURE 4
Emotional vectors plotted with regard to the third and fourth PCA components.

TABLE 1 Highest and lowest ranked words for each of the PCA dimensions.

PCA dimension	Words
PCA 1 high	together, fun, play, character, music, hate, story, interesting, amazing, album, especially, surprise, would, song, characters, interest, wish, unfortunately, perspective, love, stuff, one, happens, much, learned, ideas, quite, filled, sound, change
PCA 1 low	behavior, somehow, without, nobody, body, pain, almost, meant, happened, cause, clearly, completely, funny, away, humans, wrong, nothing, hurt, brain, others, trust, feel, saying, thinking, situation, someone, caused, truth, honestly, must
PCA 2 high	interesting, religion, seen, actually, basically, picture, crazy, different, even, clearly, wonder, literally, political, talking, beyond, individual, rather, actual, behavior, look, almost, quite, people, irony, pure, another, would, incredibly, power, exactly
PCA 2 low	pain, feel, feeling, appreciate, hear, alone, sometimes, hope, situation, life, better, good, feelings, felt, heart, true, always, everything, laugh, bad, able, thoughts, wonderful, choice, whatever, relationship, focus, anyway, loved, wish
PCA 3 high	story, happen, might, different, interesting, hope, scared, could, someone, something, weird, anyone, hear, totally, happened, never, afraid, crazy, bring, imagine, quite, would, moment, bit, alone, similar, true, surprise, nobody, together
PCA 3 low	give, good, say, literally, trying, opinion, people, words, saying, word, idea, understand, absolutely, mean, bad, play, incredibly, sound, everything, strong, either, power, behavior, move, point, every, reasons, everyone, telling, nothing
PCA 4 high	tell, imagine, moment, sad, sometimes, everyone, kinda, crying, loud, remember, little, fun, somehow, even, angry, someone, funny, feel, triggered, almost, thought, cry, still, tears, honestly, scene, seeing, hurt, scared, turn
PCA 4 low	interesting, change, give, need, opinion, anything, faith, unfortunately, means, different, situation, anyone, given, heard, deal, question, quite, rather, great, yet, something, individual, often, knowledge, hear, move, talent, nothing, however, another

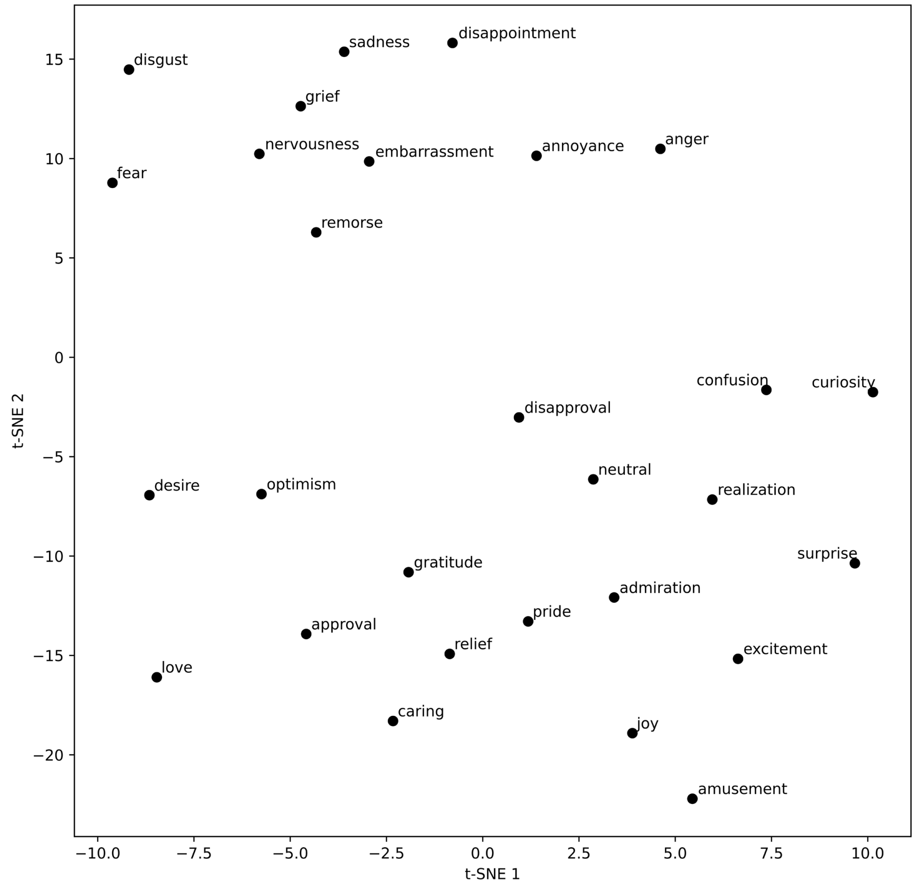


FIGURE 5
t-SNE components.

TABLE 2 Logistic regression results predicting sentiment for texts.

Variable	B	SE	z	p	95% CI
Constant	−0.401	0.259	−1.551	0.121	[−0.909, 0.106]
PCA 1	1.599	0.209	7.657	0.000	[1.189, 2.008]
PCA 2	−0.116	0.251	−0.462	0.644	[−0.609, 0.377]
PCA 3	−0.466	0.359	−1.297	0.195	[−1.170, 0.238]
PCA 4	−0.254	0.318	−0.801	0.423	[−0.877, 0.368]

Dependent variable: sentiment. Observations: 155,633. Pseudo R-squared: 0.0003071. Log-Likelihood: −103,800. LLR *p*-value: 4.684e−13.

the first two dimensions (valence and arousal). The last two dimensions were significantly less pronounced, which is most probably the effect of smaller datasets, as each of the two datasets contained only half of all the text available for the primary analysis (see Supplementary materials).

The results of the correlation tests of the placement of words alongside the different components have to be interpreted in the light of the fact that while emotion vectors synthesize the information from many documents, labeled with a given emotion, single word vectors only relate to a limited number of nearest words on each side of the specific token. This results in the word vectors carrying less information and thus not being a robust indicator of emotional expression. Still, even under these strict limitations, the first component achieves a robust correlation with its corresponding norm ($r = 0.31$; $p < 0.001$). It should be noted that the lack of significance of the other components does not prove that they are not related to their corresponding dimensions. This point is underlined by the scarcity of information embedded in their respective word vectors, as well as the fact that only a small portion of words from the GoEmotions dataset actually overlap with the available norms (Bradley and Lang, 1999).

Even though the qualitative inspection of words suffers from the same limitation of word vectors carrying less information than the emotion vectors, some interesting examples that corroborate the correlation between principal components and the dimensions of emotional experience can be found (see Table 1). Scored high on the first component (reflecting valence), are such words as together, fun, play, music, interesting, and love, all of which relate to high valence concepts. On the other side of the same component, there are words like pain, and hurt, both related to low valence concepts. For the second component, high scoring are words like interesting, crazy, power, and incredibly, which relate to high arousal; low scoring words are feeling, and alone, reflecting low arousal. The high scoring words on the third component are among others: scared, afraid, crazy, and surprise corresponding to low dominance; lower-scoring words on this component are words like absolutely, strong, power, and incredibly, reflecting high dominance (keep in mind that in the case of the third component the factor loadings are stipulated to be negatively related to the dominance dimension; see Figure 4). Finally, the words presented for the fourth component do not seem directly related to the dimension of unpredictability.

The aforementioned words mostly confirm the relation of the PCA components to the emotional dimensions; however, as can be seen from Table 1, not all of the presented words fit this pattern. Examples such as hate for high valence, laugh for low valence, wonderful, and laugh for low arousal do not fit into the outlined interpretation. These outliers could exist both due to the aforementioned problem with low informative value of specific word vectors and due to the specific ways in which they were used in their corresponding posts. Because of the high volume of the dataset, a qualitative exploration of each and every post within which they were found is impossible.

The t-SNE analysis revealed two main clusters of emotion vectors (see Figure 5). One cluster comprises negative emotions such as anger, sadness, disappointment, and remorse. The other cluster includes mainly positive emotions such as admiration, pride, excitement, joy, and amusement, as well as neutral emotions. Interestingly, disapproval, an openly non-positive emotion, is also found in this cluster. This bipolar structure confirms a significant influence of the valence dimension on the semantic arrangement of the emotion vectors. Since t-SNE focuses on preserving pairwise distances between data points (Van der Maaten and Hinton, 2008), it primarily reflects the valence dimension, while the other dimensions identified by PCA are not visible in the t-SNE visualization, as expected. Consistent with t-SNE's objective, emotions with similar meanings and expressions (e.g., desire and optimism; confusion and curiosity; sadness and disappointment) are positioned close to each other. It is important to note that the t-SNE results are sensitive to the choice of hyperparameters. In this analysis, we selected parameters that clearly delineated clusters, but different settings could produce varying results. A comprehensive exploration of all possible hyperparameters is beyond the scope of this paper.

Finally, the logistic regression results indicated that the first PCA component has a significant relationship with the sentiment of the texts ($\beta = 1.60$, $p < 0.001$; see Table 2). This finding further corroborates the conclusion that the first component reflects the valence dimension. Although the amount of variance explained by the model is very low (Pseudo R-squared: 0.0003071), this is expected because the emotion vectors used to create the PCA components were derived from the compressed information of over 50,000 texts, making it impossible to retain all information about every single text. Similar to the situation with the word vectors, the individual texts were only small snippets of the long-concatenated series that generated the emotion vectors.

4.1 Limitations

Our methodology assumes that words surrounding a specific token are indicative of its emotional connotation. However, this assumption does not consider the complexity of language and semantics. The emotional connotation of a word can significantly change depending on its position and usage in the sentence. As a result, single-word vectors may carry less information and be less reliable indicators of emotional expression. This challenge is reflected in our correlation test results, which, while statistically significant, show a relatively low correlation coefficient ($r = 0.31$). While more advanced word embedding methods that consider distant relations between words exist, such as transformer models (Vaswani et al., 2017), they are limited in the length of the text that they can represent, and thus are not sufficient for the current task where long, concatenated texts were analyzed. One possibility of using them is to average the vectors representing texts related to specific emotions, however, due to the noise inherent in this averaging, this method was not chosen for the current study.

Additionally, the interpretations of the third and fourth components of the PCA analysis might not fully correspond to the emotional dimensions of dominance and unpredictability, respectively. The third component was less varied and mainly clustered around the middle, suggesting a limited variability in dominance among the emotions. The fourth component explained the least amount of variance and its link to the dimension of unpredictability was inconclusive, especially given the unexpected positioning of certain emotions such as fear. Furthermore, there were certain word examples that did not fit the expected emotional dimensions, such as 'hate' for high valence and 'wonderful' for low arousal. While we attribute these anomalies to discourse-related artifacts and noise, they may also point to the complexity and multidimensionality of emotions that a linear component analysis may not fully capture. Another possibility points back to the information issues related to analyzing single word vectors, as they carry significantly less information than their emotion vectors counterparts.

From the methodological perspective, the fact that the emotions were labeled by the readers of text, and not their authors, stands in disagreement with the methods of previous studies, which often probed the person who experienced the emotions directly for their descriptions. One cannot expect that in all possible cases the annotator will correctly judge the emotion of the writer, or that the writer will always honestly describe their internal affairs. While the question of whether the influence of these two confounders is strong enough to produce qualitatively different results is an open one, the problem of text-based emotion communication and understanding is important in itself. This is especially true in the current age, where a lot of communication is done through text.

The preset number of emotion labels can also be seen as a limitation in the sense that by using them, the results of the current study will be biased by previous literature that has produced them. On the other hand, if annotators had been asked to describe the emotions in an open-ended manner, their results would still have to be categorized into label-like groups just the same. This grouping would be necessary to bind enough different texts together to produce robust emotion vectors. Drawing from the knowledge generated by previous studies is therefore a defensible alternative.

Finally, it is worth mentioning, that while the research on emotional components has a long history (Gendron and Feldman Barrett, 2009), the current study is to our best knowledge the first attempt at recreating emotional components based on numerical representations of natural language and, as such, is to be viewed as exploratory research. The findings of this study are best viewed as an invitation to use word embeddings to study psychological phenomena using newer, better-suited methods that allow researchers to analyze qualitative data in a quantitative manner.

4.2 Implications

Despite its exploratory nature, the current study shows that similar emotional components to the ones presented by the previous literature can be extracted from text using word embeddings. Specifically, these components were recovered by triangulating the semantic content of texts sourced from social media with peoples' judgements of what emotion the author of these texts wanted to express (limited to the 28 emotions picked for annotation). Considering the two confounders present—first the willingness of the author to honestly communicate their emotions, and second, the ability of the annotator to correctly gauge what the author wanted to communicate—it is difficult to claim that the topology reported in the current study perfectly reflects the topology of internal emotional experience. Furthermore, given that the annotators were limited in their responses to a preset list of 28 emotions based on psychological literature, this study cannot introduce novel emotional phenomena, as it is constrained to those studied by previous researchers.

However, what this study shows is that the defining dimensions of emotions, as studied through more direct, yet less ecologically valid means of questionnaires and self-reports, are reflected in the semantic structure of how they can be expressed in written language. This can be explained by the process through which our need to communicate our internal states through language shapes and creates language itself. This interpretation aligns with Chafe's work, which emphasizes that the structure and use of language are deeply influenced by the need to communicate conscious experiences and suggests that our expressions in written language naturally reflect the dimensions of internal emotional states (Chafe, 1996, 2013).

This method, when compared to the previous studies which mostly used specialized questionnaires, allows for a more ecologically valid analysis of the core dimensions of emotions. It ensures that the extracted components are grounded in the naturalistic expression of emotions and not artificially constrained by the assumptions of any particular theoretical model (Jackson et al., 2022). However, due to the indirect procurement of emotion labels (through readers and not directly from the authors), as well as the noise present in naturalistic expressions, it does not directly challenge existing methods, complementing them instead.

However, the presence of this noise could shed some light on the differences between the previous studies in the number of components that can be recovered (Bliss-Moreau et al., 2020; Fontaine et al., 2007; Mehrabian, 1996). This is evidenced by the clear dichotomy between fear and anger on the third component, supported in part by the qualitative word inspection, and by the vague sketch of unpredictability on the fourth of the recovered components. Perhaps with cleaner data and higher sample sizes, these components could be systematically recovered using classical methods. Another possibility is that laboratory studies obscure certain dimensions of emotional experience. This could be true especially for the dimension of dominance, the expression of which could be socially undesirable. Here the use of external annotators, rather than the authors of the text becomes an asset as it eradicates the influence of such social undesirability on the effects of the study.

As a last point, it is important to emphasize that the "emotion vectors" discussed in this study are purely mathematical representations derived from word embeddings, capturing the semantic and emotional content of text (Gutiérrez and Keith, 2019; Mikolov et al., 2013a,b). Unlike vectors of force in physics, which have a direction and magnitude related to physical movement, emotion vectors do not directly correspond to any physical or embodied experiences. They are abstract, numerical constructs designed to encapsulate the relationships between words in a multidimensional space, reflecting the latent structure of emotional content in language. This distinction is crucial to avoid conflating these computational representations with the physiological or psychological processes involved in action readiness, which pertains to the body's preparation for specific actions in response to emotions (Frijda, 2010). Nonetheless, this separation does not diminish the potential value of exploring how these numerical representations might correlate with or illuminate aspects of embodied emotions. Future research could delve deeper into this intersection, investigating how emotion vectors could be used to study the embodiment of emotions, perhaps by correlating these computational measures with physiological data or by incorporating word embedding techniques into previous studies that tested the influence of text data on participants' action-readiness (Lewinski et al., 2016). Such explorations could provide a richer, more integrated understanding of how emotions are represented and experienced.

4.3 Future directions

Future studies could try to recreate the current study on additional datasets of comparable quality. This would require researchers to assemble datasets of adequate length and content variance. The task of systematizing such endeavors has not been undertaken yet; however the great work done by Google (Demszky et al., 2020) can offer some directions in that regard. To our knowledge, as of yet, no dataset of comparable quality exists in open access. However, the data itself is available on the Internet, and its size is constantly growing, due to the popularity of social media sites.

Alternatively, recreating this study on a dataset with emotions annotated by the text authors instead of readers, could provide valuable information on the nature of the difference between these two emotional planes. This kind of research could shed more light on the problems related to communicating emotional information over the internet and through other text-based media, with an emphasis on the different sources of noise that partake in this process and can in many cases result in misunderstandings. The method itself can also be extended to different domains of psychology. For example, it could be well applied to the task of reconstructing the components of personality, assuming that the data are found to support this endeavor. Word embeddings can also be used in a completely data-driven way to analyze the results of qualitative interviews and create completely new psychological constructs. Furthermore, the method bypasses the difficulties in analyzing the emotional experience of individuals associated with such limitations as memory bias in answering questionnaires. The possibility of analyzing the text written by a specific individual over a span of time could therefore allow researchers to get a glimpse of what so far has been hidden behind population-wide studies—the way people express and experience emotions individually.

From a technical perspective, there is a possibility that the method of creating emotion vectors and applying PCA to them with the aim to extract emotional dimension components could be repurposed as a feature extraction method for emotion prediction. Future studies could try to apply similar techniques to this and other datasets and see whether the addition of these extracted features to more advanced machine learning models, such as deep learning architectures, XGBoost, SVM with non-linear kernels, and artificial neural networks (ANNs) leads to improved model accuracy and robustness.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/hplisiecki/emotion_topology.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the [patients/ participants OR patients/participants legal guardian/next of kin] was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

HP: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. AS: Conceptualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Polish National Center of Sciences; under Grant number [2020/38/E/HS6/00302].

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1401084/full#supplementary-material

References

Abdi, H., and Williams, L. J. (2010). Principal component analysis. *WIREs Comput. Stat.* 2, 433–459. doi: 10.1002/wics.101

Al-Amin, M., Islam, M. S., and Das Uzzal, S. (2017). Sentiment analysis of Bengali comments with Word2Vec and sentiment information of words. *2017 international conference on electrical, computer and communication engineering (ECCE)*, 186–190. doi: 10.1109/ECACE.2017.7912903

Barrett, L. F., Quigley, K. S., and Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philos. Trans. R. Soc. B, Biol. Sci.* 371:20160011. doi: 10.1098/rstb.2016.0011

Bliss-Moreau, E., Williams, L. A., and Santistevan, A. C. (2020). The immutability of valence and arousal in the foundation of emotion. *Emotion* 20, 993–1004. doi: 10.1037/emo0000606

Bradley, M. M., and Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Technical report C-1). The Center for Research in Psychophysiology, University of Florida.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., and Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vis. Res.* 41, 1179–1208. doi: 10.1016/S0042-6989(01)00002-5

Chafe, W. (1996). How consciousness shapes language. *Pragmat. Cogn.* 4, 35–54. doi: 10.1075/pc.4.1.04cha

Chafe, W. (2013). "Toward a thought-based linguistics" in Functional approaches to language. eds. S. Bischoff and C. Jany (Berlin, Germany: De Gruyter), 107–130.

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., and Banaji, M. R. (2021). Gender stereotypes in natural language: word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychol. Sci.* 32, 218–240. doi: 10.1177/0956797620963619

Cowen, A. S., and Keltner, D. (2020). What the face displays: mapping 28 emotions conveyed by naturalistic expression. *Am. Psychol.* 75, 349–364. doi: 10.1037/amp0000488

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., and Keltner, D. (2019). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nat. Hum. Behav.* 3, 369–382. doi: 10.1038/s41562-019-0533-6

Dadas, S. (2019). *Polish NLP resources* (1.0) [computer software]. Available at: https://github.com/sdadas/polish-nlp-resources (Original work published 2018)

Dellacherie, D., Bigand, E., Molin, P., Baulac, M., and Samson, S. (2011). Multidimensional scaling of emotional responses to music in patients with temporal lobe resection. *Cortex* 47, 1107–1115. doi: 10.1016/j.cortex.2011.05.007

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). *GoEmotions: a dataset of fine-grained emotions* (arXiv:2005.00547). arXiv. Available at: https://doi.org/10.48550/arXiv.2005.00547

Durrheim, K., Schuld, M., Mafunda, M., and Mazibuko, S. (2023). Using word embeddings to investigate cultural biases. *Br. J. Soc. Psychol.* 62, 617–629. doi: 10.1111/bjso.12560

Ekman, P., Davidson, R. J., and Friesen, W. V. (1990). The Duchenne smile: emotional expression and brain physiology: II. *J. Pers. Soc. Psychol.* 58, 342–353. doi: 10.1037/0022-3514.58.2.342

Evmenenko, A., and Teixeira, D. S. (2022). The circumplex model of affect in physical activity contexts: a systematic review. *Int. J. Sport Exerc. Psychol.* 20, 168–201. doi: 10.1080/1612197X.2020.1854818

Feldman, L. A. (1995). Valence focus and arousal focus: individual differences in the structure of affective experience. *J. Pers. Soc. Psychol.* 69, 153–166. doi: 10.1037/0022-3514.69.1.153

Fontaine, J. R. J., Poortinga, Y. H., Setiadi, B., and Markam, S. S. (2002). Cognitive structure of emotion terms in Indonesia and the Netherlands. *Cognit. Emot.* 16, 61–86. doi: 10.1080/02699933014000130

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychol. Sci.* 18, 1050–1057. doi: 10.1111/j.1467-9280.2007.02024.x

Frijda, N. H. (2010). Impulsive action and motivation. *Biol. Psychol.* 84, 570–579. doi: 10.1016/j.biopsycho.2010.01.005

Gendron, M., and Feldman Barrett, L. (2009). Reconstructing the past: a century of ideas about emotion in psychology. *Emot. Rev.* 1, 316–339. doi: 10.1177/1754073909338877

Gutiérrez, L., and Keith, B. (2019). "A systematic literature review on word embeddings" in Trends and applications in software engineering. eds. J. Mejia, M. Muñoz, Á. Rocha, A. Peña and M. Pérez-Cisneros (New York, United States: Springer International Publishing), 132–141.

Imbir, K. K. (2016). Affective norms for 4900 polish words reload (ANPW_R): assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Front. Psychol.* 7:1081. doi: 10.3389/fpsyg.2016.01081

Imbir, K. K., Duda-Goławska, J., Pastwa, M., Jankowska, M., Modzelewska, A., Sobieszek, A., et al. (2020). Electrophysiological and behavioral correlates of valence, arousal and subjective significance in the lexical decision task. *Front. Hum. Neurosci.* 14. doi: 10.3389/fnhum.2020.567220

Islam, M. R., Ahmmed, M. K., and Zibran, M. F. (2019). MarValous: machine learning based detection of emotions in the valence-arousal space in software engineering text. *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 1786–1793. doi: 10.1145/3297280.3297455

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., and Lindquist, K. A. (2022). From text to thought: how analyzing language can advance psychological science. *Perspect. Psychol. Sci.* 17, 805–826. doi: 10.1177/17456916211004899

Jatnika, D., Bijaksana, M. A., and Suryani, A. A. (2019). Word2Vec model analysis for semantic similarities in English words. *Procedia Comput. Sci.* 157, 160–167. doi: 10.1016/j.procs.2019.08.153

Jia, K. (2021). Chinese sentiment classification based on Word2vec and vector arithmetic in human–robot conversation. *Comput. Electr. Eng.* 95:107423. doi: 10.1016/j.compeleceng.2021.107423

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., and Rudzicz, F. (2019). A survey of word embeddings for clinical text. *J. Biomed. Inform.* 100:100057. doi: 10.1016/j.yjbinx.2019.100057

Lampier, L. C., Caldeira, E., Delisle-Rodriguez, D., Floriano, A., and Bastos-Filho, T. F. (2022). A preliminary approach to identify arousal and valence using remote Photoplethysmography. In T. F. Bastos-Filho, Oliveira CaldeiraE. M. De and A. Frizera-Neto (Eds.), XXVII Brazilian congress on biomedical engineering (Vol. *83*, pp. 1659–1664). New York, United States: Springer International Publishing

Le, Q., and Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st international conference on machine learning*, 1188–1196. Available at: https://proceedings.mlr.press/v32/le14.html (Accessed March 10, 2024).

Lewinski, P., Fransen, M. L., and Tan, E. S. (2016). Embodied resistance to persuasion in advertising. *Front. Psychol.* 7. doi: 10.3389/fpsyg.2016.01202

Lin, Y., Hoover, J., Portillo-Wightman, G., Park, C., Dehghani, M., and Ji, H. (2018). Acquiring background knowledge to improve moral value prediction. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 552–559. doi: 10.1109/ASONAM.2018.8508244

Martínez-Tejada, L. A., Maruyama, Y., Yoshimura, N., and Koike, Y. (2020). Analysis of personality and EEG features in emotion recognition using machine learning techniques to classify arousal and valence labels. *Mach. Learn. Knowl. Extr.* 2, 99–124. doi: 10.3390/make2020007

Mehrabian, A. (1996). Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in temperament. *Curr. Psychol.* 14, 261–292. doi: 10.1007/BF02686918

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv* [Preprint]. *arXiv:1301.3781.*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. Available at: https://doi.org/10.48550/ARXIV.1310.4546

Nicolaou, M. A., Gunes, H., and Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Trans. Affect. Comput.* 2, 92–105. doi: 10.1109/T-AFFC.2011.9

Nowlis, V., and Nowlis, H. H. (1956). The description and analysis of mood. *Ann. N. Y. Acad. Sci.* 65, 345–355. doi: 10.1111/j.1749-6632.1956.tb49644.x

Plisiecki, H., and Sobieszek, A. (2023). Extrapolation of affective norms using transformer-based neural networks and its application to experimental stimuli selection. *Behav. Res. Methods* 56, 4716–4731. doi: 10.3758/s13428-023-02212-3

Plutchik, R. (1980). "Chapter 1—A general psychoevolutionary theory of emotion" in Theories of emotion. eds. R. Plutchik and H. Kellerman. (Amsterdam, Netherlands: Academic Press), 3–33.

Posner, J., Russell, J. A., and Peterson, B. S. (2005). The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev. Psychopathol.* 17, 715–734. doi: 10.1017/S0954579405050340

Richie, R., Zou, W., and Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology* 5:50. doi: 10.1525/collabra.282

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178. doi: 10.1037/h0077714

Russell, J., and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *J. Res. Pers.* 11, 273–294. doi: 10.1016/0092-6566(77)90037-X

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *J. Exp. Psychol.* 44, 229–237. doi: 10.1037/h0055778

Shaver, P., Schwartz, J., Kirson, D., and O'Connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *J. Pers. Soc. Psychol.* 52, 1061–1086. doi: 10.1037/0022-3514.52.6.1061

Stanisławski, K., Cieciuch, J., and Strus, W. (2021). Ellipse rather than a circumplex: a systematic test of various circumplexes of emotions. *Personal. Individ. Differ.* 181:111052. doi: 10.1016/j.paid.2021.111052

Syssau, A., Yakhloufi, A., Giudicelli, E., Monnier, C., and Anders, R. (2021). FANCat: French affective norms for ten emotional categories. *Behav. Res. Methods* 53, 447–465. doi: 10.3758/s13428-020-01450-z

Tseng, A., Bansal, R., Liu, J., Gerber, A. J., Goh, S., Posner, J., et al. (2014). Using the circumplex model of affect to study valence and arousal ratings of emotional faces by children and adults with autism spectrum disorders. *J. Autism Dev. Disord.* 44, 1332–1346. doi: 10.1007/s10803-013-1993-6

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9:2579–2605.

Van Loon, A., and Freese, J. (2023). Word embeddings reveal how fundamental sentiments structure natural language. *Am. Behav. Sci.* 67, 175–200. doi: 10.1177/00027642211066046

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention Is All You Need. *arXiv* preprint.

Widmann, T., and Wich, M. (2022). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. *Polit. Anal.* 31, 626–641. doi: 10.1017/pan.2022.15

Wiles, J. A., and Cornwell, T. B. (1991). A review of methods utilized in measuring affect, feelings, and emotion in advertising. *Curr. Issues Res. Advert.* 13, 241–275. doi: 10.1080/01633392.1991.10504968

Woodard, K., Zettersten, M., and Pollak, S. D. (2022). The representation of emotion knowledge across development. *Child Dev.* 93, e237–e250. doi: 10.1111/cdev.13716

Yao, Z., Yu, D., Wang, L., Zhu, X., Guo, J., and Wang, Z. (2016). Effects of valence and arousal on emotional word processing are modulated by concreteness: behavioral and ERP evidence from a lexical decision task. *Int. J. Psychophysiol.* 110, 231–242. doi: 10.1016/j.ijpsycho.2016.07.499

frontiers | Frontiers in Psychology

TYPE Review
PUBLISHED 15 October 2024
DOI 10.3389/fpsyg.2024.1369957

The role of socio-emotional attributes in enhancing human-AI collaboration

Michal Kolomaznik[1]*, Vladimir Petrik[1], Michal Slama[2] and Vojtech Jurik[3,4]

[1]Czech University of Life Science, Prague, Czechia, [2]University of Hradec Kralove, Hradec Kralove, Czechia, [3]Institute of Computer Aided Engineering and Computer Science, Faculty of Civil Engineering, Brno University of Technology, Brno, Czechia, [4]Department of Psychology, Faculty of Arts, Masaryk University, Brno, Czechia

Check for updates

OPEN ACCESS

EDITED BY
Simone Belli,
Complutense University of Madrid, Spain

REVIEWED BY
Zaifeng Gao,
Zhejiang University, China
Meisam Dastani,
Gonabad University of Medical Sciences, Iran
Ayhan Dolunay,
Near East University, Cyprus

*CORRESPONDENCE
Michal Kolomaznik
✉ kolomaznik@pef.czu.cz

RECEIVED 14 April 2024
ACCEPTED 18 September 2024
PUBLISHED 15 October 2024

CITATION
Kolomaznik M, Petrik V, Slama M and Jurik V (2024) The role of socio-emotional attributes in enhancing human-AI collaboration.
Front. Psychol. 15:1369957.
doi: 10.3389/fpsyg.2024.1369957

This article delves into the dynamics of human interaction with artificial intelligence (AI), emphasizing the optimization of these interactions to enhance human productivity. Employing a Grounded Theory Literature Review (GTLR) methodology, the study systematically identifies and analyzes themes from literature published between 2018 and 2023. Data were collected primarily from the Scopus database, with the Web of Science used to corroborate findings and include additional sources identified through a snowball effect. At the heart of this exploration is the pivotal role of socio-emotional attributes such as trust, empathy, rapport, user engagement, and anthropomorphization—elements crucial for the successful integration of AI into human activities. By conducting a comprehensive review of existing literature and incorporating case studies, this study illuminates how AI systems can be designed and employed to foster deeper trust and empathetic understanding between humans and machines. The analysis reveals that when AI systems are attuned to human emotional and cognitive needs, there is a marked improvement in collaborative efficiency and productivity. Furthermore, the paper discusses the ethical implications and potential societal impacts of fostering such human-AI relationships. It argues for a paradigm shift in AI development—from focusing predominantly on technical proficiency to embracing a more holistic approach that values the socio-emotional aspects of human-AI interaction. This shift could pave the way for more meaningful and productive collaborations between humans and AI, ultimately leading to advancements that are both technologically innovative and human-centric.

KEYWORDS

autonomous technology, human–robot interaction, artificial intelligence as social actors, perception of AI, human-like AI

1 Introduction

This research explores the evolving socio-economic landscape through the lens of technological advancement and its impact on societal structures. Drawing a parallel with Edward Bellamy's visionary narrative in "Looking Backward," where protagonist Julian West wakes up to a transformed society after a 113-year slumber, this study examines similar transformative trends in contemporary societies. Bellamy's fictional account, set in 1887, presents a reimagined social structure where employment ceases at 45, succeeded by a phase of community mentorship. This societal model, emphasizing reduced working hours, facilitates personal development and community engagement, supported by comprehensive welfare systems (Bellamy and Beaumont, 2009).The current era is witnessing analogous transformative trends, primarily driven by rapid advancements in fields such as machine learning and robotics. These technological strides have significantly enhanced productivity and revolutionized various industry sectors such as finance, transportation, defense, and energy management (Brynjolfsson and McAfee, 2014; Manyika et al., 2017). Concurrently, the Internet of Things (IoT), fueled by high-speed networks and remote sensors, is facilitating unprecedented connectivity between people and businesses. Collectively, these developments hold the promise of a new era that could potentially uplift the quality of life for many individuals (West, 2018; McKinsey Global Institute, 2020).

Despite these benefits, there is a parallel and compelling narrative of apprehension and fear. That is represented in a widespread concern about the potential of AI and robotics potentially displacing jobs on a massive scale, pushing vast numbers of people into poverty, and forcing governments to consider the implementation of a universal basic income (Clifford, 2016; Stern, 2016). A study by the Pew Research Center captures this anxiety, noting that "nearly half (48%) of these experts project a future where robots and digital agents have displaced a significant proportion of both blue- and white-collar workers." Such displacement, they fear, could lead to alarming spikes in income inequality, potentially rendering large swathes of the population unemployable and triggering destabilizing effects on the social order (Smith, 2014; Frey and Osborne, 2017).

Addressing these challenges requires organizations to adapt proactively to the accelerating pace of automation, informatics, robotics, sensors, and mobile technology. This adaptation necessitates the development of change management strategies to facilitate the transition of employees into new roles that synergize with, rather than compete against, autonomous systems (Fleming et al., 2020; Davenport and Kirby, 2016). Moreover, for these autonomous systems to be effectively integrated and beneficial, it is important to ensure that people are not only capable of working with these technologies but are also inclined to do so. These systems should be perceived less as impersonal tools and more as interactive assistants, partners, or collaborators. This shift in perception, characteristic of Industry 4.0 (Schlaepfer and Koch, 2015; Schwab, 2016), is a key determinant of the successful implementation of these systems. The ability to communicate and interact effectively with these systems will be central to realizing the potential benefits of this new technological era (Ford, 2015; Kaplan, 2015; Brynjolfsson et al., 2018).

For this reason, it becomes essential to understand not only their economic and industrial impact but also their profound influence on the fabric of social interactions. This is where the study extends into exploring the role of AI in reshaping the way humans connect and communicate with each other and with technology itself (Guzman, 2018; Siau and Wang, 2018) because despite significant advancements in AI technology, existing literature largely overlooks the nuanced of AI human-like behavior in enhancing human-AI collaboration. This study uniquely contributes to the field by systematically investigating how these socio-emotional attributes can be integrated into AI systems to improve collaborative efficiency and productivity. By conducting a comprehensive review of literature and incorporating detailed case studies, this research identifies critical gaps in understanding the human-centered design of AI systems. Specifically, it addresses the need for a deeper exploration of how this human like behavior can be operationalized in AI to foster more meaningful and productive human-AI interactions.

2 AI in the world of social interactions

The transition from traditional forms of interaction to AI-mediated communication represents a significant shift in the paradigm of human relationships (Fortunati and Edwards, 2020). In this context, the philosophical insights of Martin Buber become particularly relevant. His distinction between the "I-It" and "I-Thou" relationships offers a lens through which we can examine the evolving dynamics of human interactions in an AI-augmented world. While Buber's analysis was initially focused on human-to-human relationships, the principles he outlined have newfound implications in the realm of human-AI interactions.

Central to Buber's philosophy is the idea that the essence of life is embedded within relationships. He famously stated, "Man wishes to be confirmed in his being by man and wishes to have a presence in the being of the other" (Buber and Smith, 1958). This perspective offers a unique approach through which to view AI's role in society: not merely as tools ("I-It") but as entities capable of engaging in meaningful ("I-Thou") relationships with humans to recognizing their potential role as partners in interaction.

The "I-It" approach is characterized by the perception of another human being as an object, experienced and understood predominantly through sensory impressions and external characteristics. Conversely, the "I-Thou" perspective illuminates a deeper, more intrinsic connection, acknowledging a living relationship marked by mutual recognition and profound intimacy. Buber, however, contends that such "I-Thou" encounters are not a spontaneous or natural occurrence. Rather, they demand a heightened awareness of the other's existence and an explicit shift in focus from tasks or problems to truly experiencing the partner in the interaction. He theorizes that these "I-Thou" engagements possess an unparalleled transformative potential, one that is not limited to human-human interactions but extends to connections between humans and other sentient entities (Buber, 2002).

This notion is further illustrated by the rapid advancement of communication technologies, which have transformed human interaction by removing geographical barriers and enabling collaboration independent of physical presence. The COVID-19 pandemic accelerated this digital shift, leading to a broader adoption of technologies that facilitate remote interactions (Fleming et al., 2020; Brynjolfsson et al., 2020). In this context, AI systems, like voice assistants, are evolving from being mere platforms for information exchange to becoming active participants in communication, akin to human counterparts, because of their ability to interact through various modalities—perception, expression, apparent cognition (Garnham, 1987), and communication which enriches its role in human interactions. These capabilities allow AI not just to facilitate communication but to participate in it, sometimes blurring the lines between human and machine interactions (Guzman, 2018, 2020; Zhao, 2006, p. 402; de Graaf et al., 2015,).

Indistinguishable behavior from a human partner was already presented by Alan Turing, in his paper in 1950, where he proposed an imitation game, later called the Turing test. He envisioned a future where machine interactions would become indistinguishable from human interactions, making it impossible for an observer to differentiate between the two. With recent breakthroughs in AI and robotics, we find ourselves on the cusp of this new era. Entities like game bots, robotic pets, virtual agents, and FAQ bots are integrating into daily life (Menzel and D'Aluisio, 2000; Fong et al., 2003), gradually reshaping societal norms, and increasingly approaching the threshold of passing the Turing test successfully. These advancements signify a compelling transformation in our socio-technical landscape, prompting us to revisit and reassess our notions of human-machine relationships and interactions.

3 Method

This paper uses the Grounded Theory Literature Review (GTLR) methodology, as proposed by Wolfswinkel et al. (2013), which provides a structured approach to identify prevalent themes within human-AI interaction studies. Grounded Theory, as developed by Glaser and Strauss (2017), presents an avenue for the construction of theories and the identification of thematic patterns via an inductive process encompassing data collection and analysis.

Distinct from other methodologies, Grounded Theory emphasizes an inductive orientation, contrasting the more common hypothetical-deductive perspective. This framework can serve a dual function: as a means for the generation of theoretical models emerging from data, and as a strategy for making sense of extensive data sets through coding methods (Gasson, 2009; Mattarelli et al., 2013). In the context of this paper, Grounded Theory is primarily adopted as a method for data analysis, serving to enhance the rigor in the process of identifying, selecting, and scrutinizing studies for review.

Essentially, the GTLR method treats the content encapsulated within the reviewed articles as empirical data, subjected to analysis for theme development. This approach has found utility in numerous systematic reviews in the Human-Computer Interaction (HCI) discipline (Nunes et al., 2015; Mencarini et al., 2019) and consists of four distinct stages:

(i) Define: This stage includes the determination of the inclusion/exclusion criteria, the identification of appropriate data sources, and the formulation of the specific search query.
(ii) Search: This phase encompasses the collection of articles from all the determined sources.
(iii) Select: This stage involves the establishment of the final sample by cross-referencing the gathered papers with the predetermined inclusion/exclusion criteria.
(iv) Analyze: At this stage, the chosen papers are subjected to analysis using open, axial, and selective coding techniques.

Subsequently, the presentation and discussion of the analyzed papers represent an additional stage of the methodology.

3.1 Inclusion and Exclusion criteria

3.1.1 Inclusion criteria

In accordance with the review methodology employed for this research, the object of focus was a literature review of available materials within specified period and key words. This specification aimed to curtail potential bias stemming from overlapping data within different reports from the same investigation. The inclusion criteria were then established as follows:

i **Relevance**: Research papers must primarily concern interactions between humans and AI, specifically relating to the integration of human-like behavior in AI systems and their impact on human-AI collaboration.
ii **User-centric focus**: Each included article should contain at least one user-centric study. This stipulation was imposed to ensure that the focus remained on the interaction between users and the AI system rather than on the technological performance.
iii **Human dimensions**: The papers selected must delve into the nature of the interaction between the user and the AI, with an emphasis on the human dimensions of this interaction. They should present insights on user experiences during these interactions, for example, the user's emotional response, behavior, cognitive processes, perceptions of the AI, and anticipations or evaluations of the interaction. Although the selection was not strictly limited to papers from the HCI field, the emphasis on human-centric interaction ensured relevance to the HCI community.
iv **Publication quality**: The papers should be published in peer-reviewed international journals in the final stage of publication.
v **Recency**: The timeframe of publication was set from January 2018 to December 2023 to ensure the inclusion of research conducted in an era where interaction with AI technology was not completely novel to users. While the major surge in AI usage can be attributed to the year 2014 (Grudin and Jacques, 2019), the inclusion of early 2012s research permitted an exploration of user experiences during the phase when conversational agent technology was starting to permeate public awareness.
vi **Diversity**: To provide a broad perspective, we selected studies from various industries, including healthcare, education, customer service, and finance.
vii **Empirical evidence**: We prioritized papers that provided empirical data and detailed descriptions of AI implementations, user interactions, and outcomes.

3.1.2 Exclusion criteria

In light of the established inclusion criteria, several studies were necessarily excluded from the review. Studies primarily concerned with evaluating the efficacy of AI in carrying out specific tasks without considering the user's interaction experience.

For instance, investigations into AI-enabled augmented reality (e.g., Sabeti et al., 2021) that centered solely on design framework in delivering appropriate recommendations for developers and therefore were not included. Similarly, research on teaching AI agents to understand and generate contextually relevant natural language with a goal-oriented approach (e.g., Ammanabrolu and Riedl, 2021), which focused only on the success of interaction (i.e., machine learning), without addressing aspects of user interaction, were also left out.

Studies that deployed user testing solely to evaluate the efficiency of a specific Natural Language Processing (NLP) technology or algorithm (for example, assessing an algorithm's aptitude for classifying users' intentions) were excluded as well. Research papers concerning Embodied Conversational Agents (ECAs), speech technology, or AI incapable of maintaining substantial conversation were not considered.

In addition, papers that were part of the supplementary proceedings of conferences (such as posters, workshop papers), or chapters in books, were also deemed outside the scope of this review.

3.1.3 Search approach

The review was conducted in January 2024, utilizing Scopus as the databases for sourcing relevant scholarly articles on interactions between humans and AI (Littell et al., 2008). These databases were chosen due to their extensive breadth of content, which ensures a comprehensive coverage of critical topics related to human-AI interaction. Scopus served as the primary source of data, while Web of Science was used to corroborate findings and identify additional sources through a snowball effect. This approach allowed for a thorough and inclusive review of the literature, ensuring that a wide array of perspectives and studies were considered in the analysis.

The search strategy employed a combination of terms and connectors aimed at capturing a broad spectrum of studies on AI, particularly conversational assistants, and their interaction with humans. Terminologies utilized in the search queries were selected to encompass the varying nuances of human-AI interaction, ensuring a comprehensive capture of the phenomenon from multiple perspectives.

The lexicon for the search terms was constructed iteratively to the refinement process. The intention was to emulate the best practices used in similar reviews within the HCI research field (for instance, ter Stal et al., 2020), ensuring a rigorous, yet broad coverage of relevant literature in the domain of human-AI interaction. The initial query were based on authors experience and contained "Human-AI Interaction" and "Human Factors in AI." This resulted in significant amount of papers. Initial analyses of all papers then led to the following query determined the final set of articles:

The search in Scopus was specifically constrained to the title, abstract, and keywords sections. The types of documents included in the search were primarily articles published in a journal. The chosen timeline for the search spanned from January 2018 to December 2023, in order to capture a substantial yet manageable body of literature.

Upon execution of the search strategy, 337 entries were procured from Scopus. These results were exported to a table and harmonized for consistency. A preliminary evaluation was then undertaken to exclude any papers that evidently did not meet the pre-established criteria. This distillation process resulted in a final count of 108 papers deemed suitable for comprehensive analysis. This set of papers provided a substantive insight into human-AI interactions, yielding a critical understanding of the domain under investigation.

3.1.4 Data analysis

The selected set of 108 articles were thoroughly examined by leveraging the core principles of Grounded Theory. The main objective of this process was to discern and highlight recurring themes within the selected literature. At the start of the analysis, one or more conceptual labels were ascribed to each article, reflecting the key ideas, patterns, and insights perceived. In the subsequent phase of axial coding, these discrete codes were grouped into broader conceptual categories.

The final stage of selective coding saw the authors collectively discuss and reconcile discrepancies in the axial coding results. This stage was instrumental in weaving the individual categories into an integrated and coherent explanatory scheme.

3.1.5 Paper elimination and validation

The initial selection of articles underwent a preliminary screening based on the titles and abstracts, aligning with the set eligibility criteria. This screening led to the exclusion of 56 papers due to various reasons, such as non-alignment with the subject matter, a publication type, access. and lack of essential data (Figure 1).

Subsequently, the remaining 281 papers deemed potentially eligible were subjected to abstract review and later to a comprehensive evaluation of their full texts. This in-depth assessment was undertaken to ensure strict adherence to the inclusion criteria and research objectives Wolfswinke. The culmination of this rigorous evaluation, a total of 65 papers were chosen for the analyses as they offer a wide range of perspectives and insights on human-AI interaction, aligning with the research's main objectives.

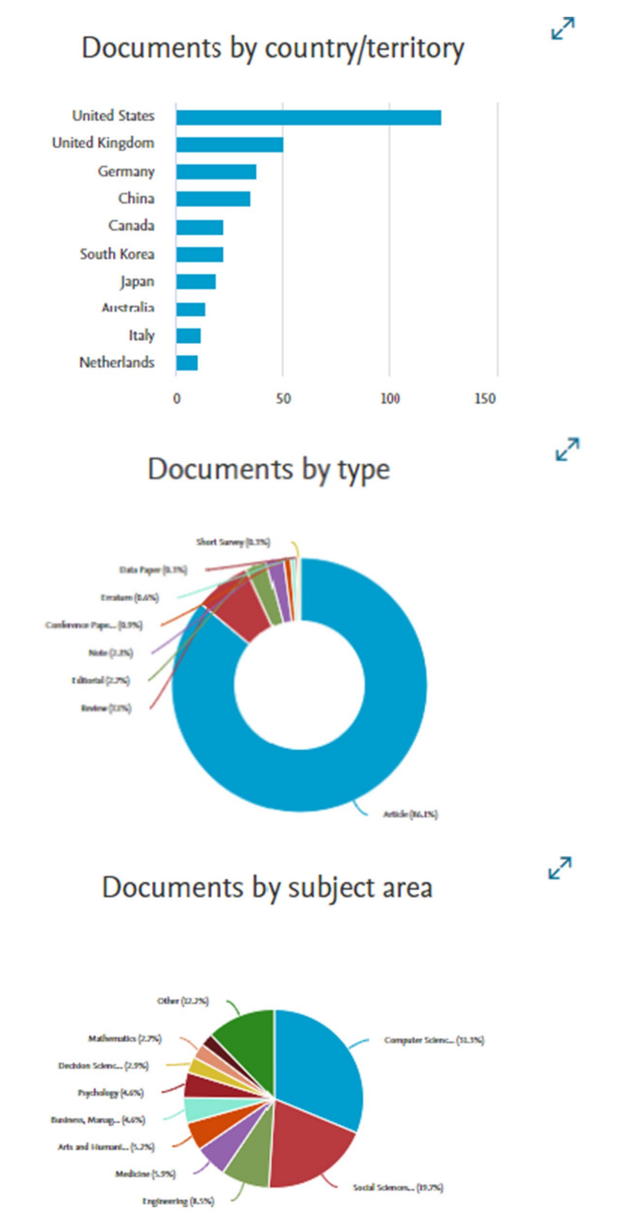


FIGURE 1
Representation of search results.

In the final stage, snowballing was used, screening the references cited within the included articles, using a similar method as applied in the previous stages of the database searches. This additional layer of screening resulted in the identification of 44 more papers, thereby expanding the corpus to include a total of 108 papers.

Figure 2 provides a detailed visual representation of the article selection process, illustrating the stages of database searches, screening, and the final count of included articles.

3.1.6 Paper analyses

In the "analyze" stage, a systematic approach was employed to extract key themes from the selected papers, adhering to the principles of Grounded Theory. The process involved several essential steps: open coding, axial coding, and selective coding. The following details the methodology used:

1 **Initial labeling**:

- **Initial review**: Each paper was thoroughly examined to identify significant ideas and observations related to human-AI interaction. Descriptive labels (codes) were assigned to capture the essence of these key points. For instance, discussions concerning "trust in AI" were coded as "Trust," while references to "AI's ability to understand emotions" were labeled "Empathy."
- **Consistency**: Although codes were derived inductively from the data, consistency was ensured by developing a coding guide. This guide included definitions and criteria for each code, enabling uniform application across similar concepts found in different papers.

2 **Axial coding**:

- **Grouping labels into categories**: Following the initial coding, the codes were reviewed to identify patterns and relationships among them. This step involved organizing the codes into broader categories. For example, codes related to "Trust," "Transparency," and "Reliability" were grouped under the larger category of "Trust Factors."
- **Exploring relationships**: Relationships between these categories were then examined to build a more integrated understanding of the socio-emotional aspects of human-AI interaction.

3 **Selective coding**:

- **Theme development**: In this final stage, the categories were synthesized into core themes that encapsulated the key insights from the analysis. The goal was to identify central themes that accurately represented the data. Themes were refined through multiple rounds of review to ensure they reflected the content of the papers effectively.
- **Validation and consensus**: The coding process included collaborative discussions to resolve any discrepancies, ensuring that the themes were robust and well-supported by the data.

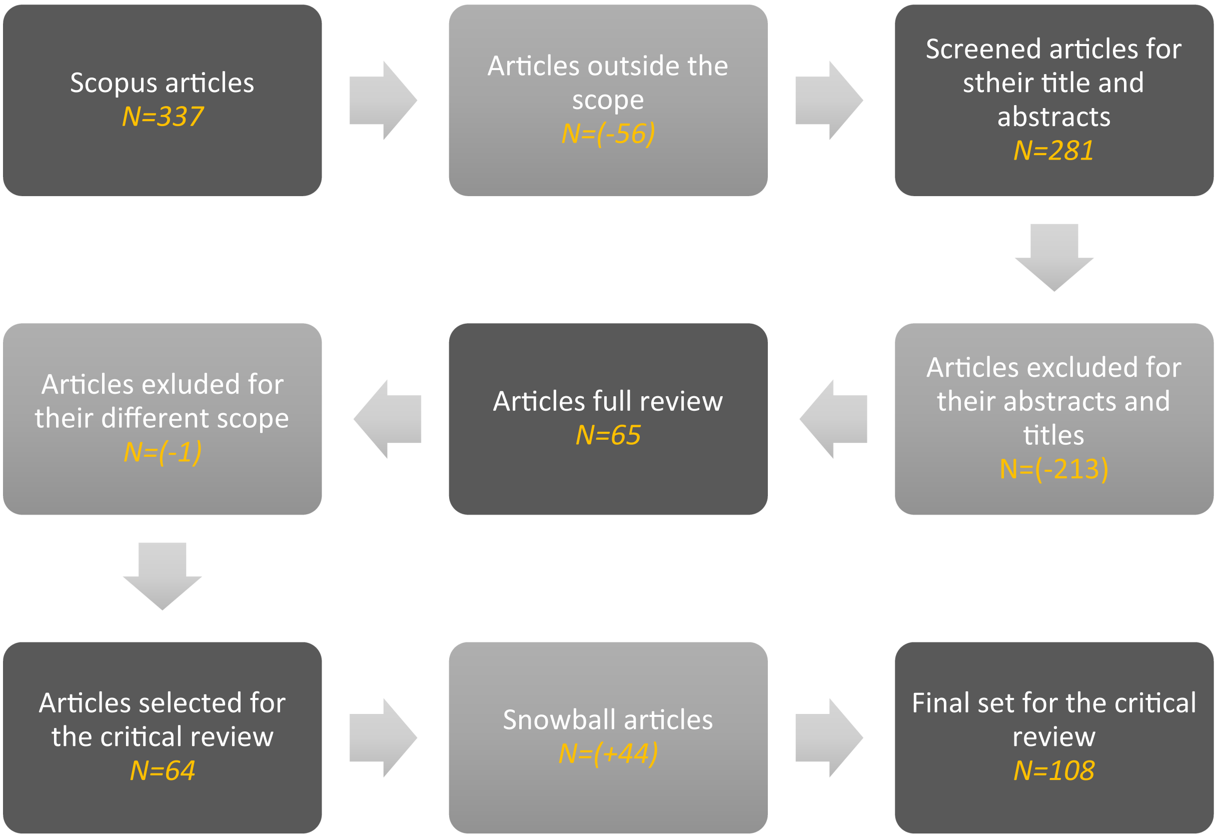

Scopus articles
N=337
Articles outside the scope
N=(-56)
Screened articles for stheir title and abstracts
N=281
Articles excluded for their abstracts and titles
N=(-213)
Articles full review
N=65
Articles exluded for their different scope
N=(-1)
Articles selected for the critical review
N=64
Snowball articles
N=(+44)
Final set for the critical review
N=108

FIGURE 2
Article selection process.

4 **Theme extraction**:

- **Final themes**: This process resulted in the identification of five key themes—Rapport, Trust, User engagement, Empathy, and Anthropomorphization. These themes were derived from recurring patterns and significant connections observed during the coding process (Table 1).

4 Discussion

4.1 Human-like perception of AI

While pre-programmed robots do not require human interaction and run independently on their human partners to deliver the value, social robots designed to communicate on a deeper level with their human opponents require some effort from their collaborators to create a meaningful outcome. Those robots are the ones that can replicate a variety of signs leading to a feeling of a social appearance (Lombard and Xu, 2021) as captured in Table 2. Such perception comes from the brain, which processes data based on experience and translates them to information, also known as the mindless process, due to three aspects (Nass and Moon, 2000): (a) people rely on their previous experience from human interactions even though they are interacting with a robot, (b) apply known social norms, (c) involve the System 1 (Kahneman, 2011). In other words, if an artificial intelligence provides sufficient cues associated with humaneness, our System 1 applies a learned script for human interaction to the interaction with a robot. Once that happens, humans do not look for more cues anymore, potentially triggering a different response. They are saturated.

These ingrained scripts were evolutionarily beneficial in our survival within the natural world (Kahneman, 2011). However, there were instances when they led to grave errors (Tversky and Kahneman, 1974), a principle that also holds relevance in the realm of artificial intelligence. Humans can be easily misled by AI cues, led to believe that they are interacting with sentient beings, thereby applying learned norms and scripts. This phenomenon was starkly illustrated by a recent incident where a finance worker at a multinational firm in Hong Kong was deceived into transferring $25 million to fraudsters. The scammers used deepfake technology to impersonate the company’s chief financial officer during a video conference call (The Register, 2024). This example underscores the inherent human tendency to trust rather than doubt, as skepticism requires more cognitive effort (Gilbert, 1991). Kahneman (2011) elaborates on this by describing two cognitive systems that influence our perception: one that is fast, effortless, and comfortable, and another that is slow, energy-intensive, and challenging to utilize. Lee (2004) further substantiates this, suggesting that our default cognitive process is inclined toward belief unless we encounter compelling evidence to the contrary. These inclinations might elucidate why people relate to AI in the same manner they relate to other humans.

If an AI system can successfully activate our social scripts, humans would instinctively respond with their ingrained social responses. This recognition of something familiar suggests that a few AI cues can elicit a perception of social interaction, prompting us to employ our automated system of social scripts (Kahneman, 2011). Simply put, if AI can stimulate humans through specific cues to trigger these learned social scripts, humans will interact with technology in the same way they interact with other people. Consequently, comprehending these cues becomes imperative for developing effective adoption frameworks, ensuring that technology is utilized to its fullest potential.

TABLE 1 Search query.

“Human-AI Interaction” OR “User Experience with AI” OR “Human-AI Collaboration” OR “Human Factors in AI” OR “AI User Perception” OR “AI Trust and Transparency” OR “Emotional AI” OR “AI Ethics in Human Interaction” OR “AI Personalization and Adaptivity”

In synthesizing the findings from the comprehensive set of articles reviewed, a range of attributes emerge that characterize AI as human-like. These attributes include Rapport, Empathy, Trust, User engagment, Anthropomorphization, and Communication. Additionally, attributes such as robust Social Interactions; a sense of Self-efficacy; expressive Body Language; capabilities for Self-prolongation or Self-preservation; fostering Friendship & Companionship; Personalization; Intuition; Creativity; respect for Privacy and Non-judgmental interaction; adherence to Ethics; logical Reasoning; the ability to Surprise and demonstrate Unpredictability; Adaptivity; Autonomy; Co-Creativity and Complementing human efforts; Competence; a distinct Identity; Memory retention; and being Culturally and Socially aware, are all identified as key factors. The subsequent sections delve into a detailed exploration of the five most prevalent themes, underscoring their significance and implications in human-AI interaction.

These key themes are:

a **Rapport** which is conceptualized as a harmonious relationship underpinned by mutual understanding and empathetic engagement between interacting entities (Gremler and Gwinner, 2008).

b **Trust** which represents a critical evaluative construct encompassing both cognitive and emotional dimensions. It reflects the user’s confidence in the reliability, integrity, and competence of the AI system (Mayer et al., 1995).

c **User engagement** in the realm of human-AI interaction denotes the user’s emotional investment and sustained engagement with the AI system. It is influenced by the perceived usefulness and satisfaction derived from the interaction (O’Brien and Toms, 2008).

d **Empathy** defined as the ability to understand and share the feelings of another. In AI interactions, empathy involves the recognition and appropriate response to user emotions (Mehrabian and Epstein, 1972).

e **Anthropomorphization** which refers to the attribution of human characteristics, behaviors, and emotions to non-human entities, such as AI (Lombard and Ditton, 1997; Davenport et al., 2020). This process enhances user acceptance and satisfaction by making AI appear more relatable and engaging (Xiao and Kumar, 2021).

4.2 Human-like attributes

4.2.1 Rapport

The recent development of AI, mainly effective advanced online chatbots, can provide those cues to build bonds with their human partners and, as such, to pass the Turing test in several restricted areas (Goodwins, 2001). Personified agents adapted to remembering the history of our discussions and advanced in imitating non-verbal

TABLE 2 Sample of anthropomorphised AI services.

Service example	Industry	Description	Anthropomorphic features	Launch
Amelia	Banking, healthcare, insurance, and telecommunications.	Amelia is an AI platform designed to automate business processes that would typically require human intelligence. It's capable of learning, expressing emotions, understanding context, and solving complex problems.	It is constructed to mimic human cognitive processes in understanding, learning, and engagement, enabling her to grasp natural language, maintain context continuity throughout conversations, manage complex inquiries, and evolve from past interactions. Amelia's capacity to process emotional nuances both in terms of comprehension and expression imparts a dimension of human-like interaction that enhances user engagement. Amelia is also capable of detecting subtle indications in the user's language to discern their emotional state and modulate her responses accordingly.	2014
Siri	Consumer technology	Siri is Apple's voice-activated virtual assistant, available on iOS devices, and capable of setting reminders, sending texts, answering questions, and other tasks.	Siri is crafted to comprehend natural human language and can even respond to more intricate, context-dependent commands. Siri also exhibits elements of personality, incorporating humor in her responses, and can adapt to the individual language usage and preferences of users over time	2010
Cortana	Consumer technology	Cortana is Microsoft's virtual assistant, integrated into Windows devices, which can set reminders, recognize natural voice, answer questions using information from Bing, etc.	Cortana's AI platform integrates chatbot services that can emulate human conversation patterns. The Language Understanding Intelligent Service (LUIS) by Microsoft is specifically engineered to comprehend and interpret human language.	2014
Alexa	Consumer technology	Alexa is Amazon's voice-activated virtual assistant, found on Echo devices, which can answer questions, play music, control smart home devices, and more.	Alexa exemplifies an anthropomorphized form of AI, demonstrating proficiency in engaging in human-like conversation and comprehending context within dialogs.	2014
Duolingo	Education	Duolingo is a language-learning platform that uses AI to adapt to users' learning habits and provide personalized education paths.	Duolingo's AI is personified as an amiable anthropomorphic owl named Duo. This owl proffers encouragement, reminders, and celebratory messages, fostering a more immersive learning environment. Duo's interactions with users further imbue the application with a sense of human-like presence.	2011
SoundHound	Music and entertainment	SoundHound is an app that can identify songs playing around you. It also offers voice-recognition features, allowing users to conduct searches or control playback with voice commands.	SoundHound displays proficiency in comprehending natural human language and contextual cues, thus exhibiting anthropomorphic characteristics in its interactions.	2009
Genesis Toys—My Friend Cayla Doll	Toy	Genesis toy is an interactive doll that uses speech recognition to converse with children, answer questions, and tell stories.	Cayla, equipped to comprehend and respond to user's speech, answer questions, and even tell stories, mirrors the interactive capabilities of a human friend. The physical design of the doll, coupled with its interactive capabilities, significantly amplifies its anthropomorphic nature.	2014
Woebot	Healthcare	Woebot is an AI-powered chatbot designed to help users manage their mental health. It uses principles of cognitive-behavioral therapy to offer guidance and support.	Woebot employs a conversational tone and expresses empathy, thus portraying human-like characteristics.	2017
Salesforce Einstein	Business	Einstein is an AI layer in the Salesforce platform that uses machine learning to predict outcomes, recommend next steps, automate tasks, and analyze data.	Einstein has the capacity to comprehend and anticipate user behavior in a manner analogous to human anticipation.	2016
ChatGPT	No limits	ChatGPT is a language model developed by OpenAI. It uses machine learning to generate human-like text based on the prompts it's given.	As an interactive language model, ChatGPT is capable of emulating human interaction and comprehending context, reflecting human-like interaction characteristics.	2020

(Continued)

TABLE 2 (Continued)

Service example	Industry	Description	Anthropomorphic features	Launch
Dali	Digital media	Dali is a large transformer model trained by OpenAI, capable of generating images from textual descriptions, displaying creativity and a high degree of abstraction.	Dali is engineered to generate images from textual descriptions, a process that requires a high degree of abstraction and creativity - traits typically associated with human intelligence.	2021
Adobe Sensei	Digital media	Adobe Sensei is an AI and machine learning framework that powers intelligent features across Adobe's products, helping users with tasks like auto-tagging photos, optimizing marketing campaigns, and more.	Sensei is engineered to emulate human perception within its image recognition capabilities. Furthermore, its automation of tasks might give an impression of 'understanding' user's requirements, reflecting human-like perceptual skills.	2016

communication are being introduced into new areas previously unimaginable. Stronks et al. (2002) reported that AI humanoids capable of speaking multiple languages and recalling past interactions significantly improved user satisfaction, with 85% of participants indicating a stronger sense of connection and rapport. *Inter alia*, those humanoids that already entered households as pets similar to Aibo, intelligent assistants like Alexa, or intimate dolls reaching new levels of sensual relationships demonstrate co-living principles and a delicate attachment to their owners (Knafo, 2015). Those Androids are not yet in mass production. However, its research indicates that in experimental circumstances, they are able to speak a variety of languages, remember previous decisions, cook and clean, entertain young kids, or as mentioned above, become an intimate companion or a companion for elderly people to cope with their loneliness, etc. till 2035.

To unlock their full potential, AI systems need to establish rapport with their human counterparts, facilitating harmonious relationships rooted in mutual understanding of feelings or ideas. This rapport could be enhanced by factors such as sensitivity and humor (Niculescu et al., 2013), which increase likability and foster cooperative activity by 30% (Short et al., 2010; Argyle, 1990). Previous studies have shown the positive impact of rapport on team effectiveness, satisfaction, and overall well-being (Morrison, 2009). Thus, the subsequent sections of this review study will delve deeper into the foundational principles of rapport necessary for successful outcomes.

Rapport, as previously discussed, is a harmonious relationship underpinned by effective communication. It rests on mutual understanding and shared experiences between human beings (Ädel, 2011), and is a synergistic process amplified by reciprocation (Ädel, 2011; Gremler and Gwinner, 2008).

A vital aspect of building rapport is the identification and demonstration of commonalities between the interacting parties (Gremler and Gwinner, 2008). This process begins with the initial affiliation and continues throughout the interaction. For example, both parties may discuss topics outside of the main subject of conversation, often involving aspects of their social or private lives, indicating an increase in trust (Ädel, 2011). Familiarity can be enhanced by tapping into shared memories, vocabulary, or knowledge (Argyle, 1990). Other methods to strengthen rapport include the use of inclusive language ("we") to foster a sense of community (Driskell et al., 2013), or mimicking the behaviors of the other party (Gremler and Gwinner, 2008).

Attributes of rapport also encompass positivity and friendliness, typically manifested through cheerfulness, praise, and enthusiasm (Ädel, 2011). Demonstrating empathy (Gremler and Gwinner, 2008) and active listening can evoke a sense of importance in the other party. Body language and verbal assurance [e.g., "hmm," "I see," etc., Gremler and Gwinner (2008)] also play critical roles. Even in challenging situations, respectful responses, such as offering an apology, can contribute to rapport building.

4.2.2 Empathy

The empathy expression by AI can significantly alter the interaction quality, particularly regarding engagement and relationship cultivation (Vossen et al., 2015; Mehrabian and Epstein, 1972). Liu and Sundar (2018) posit that an AI display of affective empathy, when consulted for health advice, can come across as more supportive than simply relaying medical data. For instance, their research demonstrated that participants who interacted with an empathetic AI were 20% more likely to follow the health advice provided, indicating a substantial increase in perceived support and trust. In line with this, Fitzpatrick et al. (2017) devised Woebot, a self-help AI for college students experiencing a 22% reduction in anxiety symptoms over two weeks,. It was found that users appreciated the AI's empathetic responses, hinting at the possibility of establishing therapeutic relationships with nonhuman agents, as long as they can express empathy. In a similar vein, Ta et al. (2020) proposed that chatbots may serve as daily companions, offering emotional support and enhancing positive emotions. Their examination of Replika user reviews and questionnaire responses emphasized that 74% of users felt that the AI's expressions of care, love, and empathy significantly improved their mood and provided a sense of companionship. Furthermore, Portela and Granell-Canut (2017) explored the impact of empathetic responses from AI on user interaction. Their findings indicated that 68% of users reported feeling more emotionally engaged when the AI expressed empathy, compared to interactions with non-empathetic AI On the hand, AI can sometimes irritate users when attempting to imitate human behavior (Urakami et al., 2019).

4.2.3 Trust

Trust, in its essence, represents a cognitive evaluation heavily influenced by both rational judgment and the emotional satisfaction connected to the feeling of security about 35% if AI offers transparent explanations for its actions, compared to the one that did not (Mayer et al., 1995; Frison et al., 2019). There is a discernable connection between User Experience (UX) and trust, as shown in UX-related studies. In the context of interactions between humans and AI, trust takes on a significant role, especially when the decisions made by the AI have considerable repercussions for the end users (Zamora, 2017). Efforts have been made to unravel the elements that can sway a user's trust during interactions with an AI. A notable study by Følstad et al. (2018) revealed that the capacity of AI to comprehend and respond appropriately to user requests, embody human-like attributes, and effectively showcase their capabilities can significantly enhance user trust by 40%. Additionally, factors such as the brand reputation and the clear communication of security and privacy measures can influence how users perceive trust. This requirement for trust is also affirmed when there are high-risk data and privacy considerations involved in the interaction with AI (Zamora, 2017). A research conducted by Yen and Chiang (2020), using data from 204 questionnaires, underscored that the perceived trustworthiness of AI is shaped by their credibility, competency, human-likeness, presence, and the quality of information they provide.

4.2.4 User engagement

In the context of user experience, engagement embodies a complex construct that integrates affective, cognitive, and behavioral interactions with technology, leading to complete absorption in the activity at hand (O'Brien and Toms, 2008; Attfield et al., 2011; Ren, 2016; Goethe et al., 2019). It encapsulates subjective experiences such as immersion, participation, and pleasure, instrumental in driving sustained user commitment (Brown and Cairns, 2004; Peters et al., 2009; Boyle et al., 2012; Saket et al., 2016; Liu et al., 2017; Lukoff et al., 2018; Debeauvais, 2016.)

Variety of factors contribute to user engagement. Prolonged interactions and heightened message interactivity with chatbots, for example, have been shown to intensify user engagement by 35% (Sundar et al., 2016; Cervone et al., 2018). Moreover, elements like emojis usage, effective listening capabilities, and prompt responses have been observed to bolster user interaction levels (Avula et al., 2018; Fadhil et al., 2018; Xiao et al., 2020). However, Ruan et al. (2019) highlighted a potential conflict between engagement and effectiveness, indicating that while entertaining and interactive AI-facilitated learning experiences were favored, they might inadvertently lead to learning inefficiencies. Their study suggested that while 60% of users enjoyed interactive learning experiences, only 45% found them effective in achieving their learning goals. Consequently, the creation of engaging AI experiences necessitates a harmonious interplay between effectiveness and engagement.

4.2.5 Anthropomorphization

In the exploration of anthropomorphization within artificial intelligence (AI) applications, research indicates a key capability for AI to imitate human intelligence traits (Syam and Sharma, 2018). Such imitation, facilitated by technological advancements in machine learning, natural language processing, and image recognition, has been seen to enhance user acceptance and satisfaction (Xiao and Kumar, 2021; Sheehan et al., 2020). Furthermore, the significance of high degrees of anthropomorphization has been linked to improved assessments of robots' social cognition (Yoganathan et al., 2021). Despite these strides, it is noteworthy that most current models only distinguish between high and low degrees of anthropomorphization, with little attention to how various types, such as physical, personality or emotional anthropomorphism, might enhance these outcomes (Davenport et al., 2020). There is a gap in the literature on how users interact with these anthropomorphized agents from the viewpoint of their self-concept (MacInnis and Folkes, 2017).

While the aforementioned studies primarily examine anthropomorphized robots and embodied conversational agents, the interaction between humans and AI has also received attention. In this context, a key aspect of research has been the perception of humanness in AI, particularly how this perception impacts the user experience (Ho and MacDorman, 2017). For instance, Schwind et al. (2018) found that AI systems with physical anthropomorphism, such as avatars with human-like faces and body language, were rated 18% higher in terms of user likability and engagement. Other factors such as language use, the ability to exhibit humor, and error occurrence have been found to influence perceived humanness (Westerman et al., 2019; Araujo, 2018). Despite that, the preference for human-like conversations is context-dependent, and not all human-like features are favored in every setting (Svenningsson and Faraon, 2019). As such, the dimensions of naturalness in AI conversations, such as conscientiousness and originality, warrant further exploration (Morrissey and Kirakowski, 2013).

Research concerning the human-like characteristics of AI and their effects on users has revealed intriguing and sometimes contrasting findings. The uncanny valley theory proposes that when non-human agents appear almost but not entirely human, they can trigger unease or even repulsion among human observers. A number of studies (e.g., Stein and Ohler, 2017; Liu and Sundar, 2018) have examined this concept in relation to AI, with mixed results. While some studies found that human-like chatbots can evoke feelings of unease, others (e.g., Skjuve et al., 2019; Ciechanowski et al., 2017) observed no such effect.

4.3 Interrelationships between themes

The intricate relationships between these themes elucidate the socio-emotional dynamics in human-AI interaction

- **Rapport and trust**: Rapport and trust are closely connected in human-AI interactions. Establishing rapport through mutual understanding and effective communication builds trust, as users feel more confident in the AI's reliability and empathy (Gremler and Gwinner, 2008; Mayer et al., 1995). This trust is essential for a secure and positive user experience (Følstad et al., 2018).
- **Rapport and user engagement**: Rapport directly influences user engagement by making interactions more meaningful and enjoyable. As users feel understood and valued by the AI, their level of engagement increases, creating a reinforcing cycle that enhances the overall experience (Avula et al., 2018; Liu et al., 2017).
- **Rapport and empathy**: Empathy is a critical factor in building rapport. AI systems that effectively recognize and respond to user emotions foster deeper mutual understanding and emotional alignment, thereby strengthening rapport (Gremler and Gwinner, 2008).
- **Rapport and anthropomorphization**: The development of rapport is facilitated by anthropomorphization. Human-like features in AI, such as humor and sensitivity, contribute to a sense of connection and mutual understanding, which are essential components of rapport (Niculescu et al., 2013).
- **Trust and user engagement**: Trust serves as a foundational element for user engagement. A high level of trust in the AI system reduces perceived risks and enhances the user's confidence, leading

to greater emotional investment and sustained engagement (Følstad et al., 2018; Yen and Chiang, 2020).

- **Trust and user engagement**: Trust is a crucial driver of user engagement in human-AI interactions. When users trust an AI system, they are more likely to engage deeply, as trust reduces perceived risks and enhances confidence in the system's reliability and functionality (Følstad et al., 2018; Zamora, 2017; Siau and Wang, 2018). Trust can also amplify the willingness to explore and utilize more features of the AI, leading to sustained and meaningful interactions (Avula et al., 2018; Sundar et al., 2016). In contexts where trust is established, users are more inclined to immerse themselves in the experience, resulting in higher levels of engagement (Boyle et al., 2012).
- **Trust and empathy**: Trust and empathy are deeply interconnected in fostering positive human-AI relationships. An AI system that effectively demonstrates empathy can enhance user trust by signaling that it not only understands the user's emotional state but also responds appropriately to it (Liu and Sundar, 2018; Mehrabian and Epstein, 1972; Fitzpatrick et al., 2017). This empathetic interaction creates a perception of the AI as being supportive and considerate, which strengthens trust (Portela and Granell-Canut, 2017). Moreover, the ability of AI to exhibit empathy can bridge the gap between human and machine, making users feel safer and more understood, thereby reinforcing their trust (Bickmore and Picard, 2005).
- **User engagement and empathy**: Empathy enhances user engagement by creating a more personalized and emotionally resonant interaction. When AI systems respond to users' emotions, they foster a deeper connection, leading to increased and sustained engagement (Liu and Sundar, 2018; Fitzpatrick et al., 2017; Bickmore and Picard, 2005).
- **User engagement and anthropomorphization**: Anthropomorphization boosts user engagement by making AI systems more relatable and human-like. When AI mimics human behaviors, users are more likely to interact with it naturally, leading to a more immersive and engaging experience (Xiao and Kumar, 2021; Sheehan et al., 2020; Nass and Moon, 2000).
- **Anthropomorphization and empathy**: Anthropomorphization aids in the expression of empathy by endowing AI with human-like qualities that facilitate emotional recognition and appropriate responses. This enhances the perceived empathy of AI systems, contributing to higher user satisfaction (Liu and Sundar, 2018).

The interrelationships among rapport, trust, user engagement, empathy, and anthropomorphization reveal the intricate socio-emotional dynamics that are foundational to human-AI interactions. These interconnected themes significantly enrich the user experience, highlighting the imperative for AI systems to be designed with a comprehensive understanding of human emotional and cognitive processes. By integrating these socio-emotional elements, AI systems can more effectively resonate with users, fostering deeper and more meaningful engagements.

5 Ethical considerations

The integration of AI in human interactions necessitates addressing ethical considerations through established frameworks such as deontological ethics, utilitarianism, and virtue ethics. Deontological ethics emphasize adherence to rules and duties, highlighting the need for AI systems to comply with ethical guidelines to ensure transparency and respect for user privacy (Floridi and Cowls, 2019; Jobin et al., 2019). Utilitarianism, which evaluates the morality of actions based on their outcomes, calls for a balance between the benefits of AI-enhanced productivity and the potential risks, such as job displacement and over-reliance on AI for emotional support (Brynjolfsson and McAfee, 2014; Bostrom and Yudkowsky, 2014). Virtue ethics focuses on developing AI systems that embody moral virtues like honesty, empathy, and integrity, promoting ethical behavior in interactions (Coeckelbergh, 2010; Turkle, 2011). Practical guidelines derived from these frameworks include designing AI with transparency, prioritizing data protection, promoting positive social interactions, and conducting continuous ethical assessments. Incorporating these ethical considerations ensures that AI systems enhance productivity while upholding the highest ethical standards, contributing to a just and equitable society (Floridi et al., 2018; Moor, 2006).

6 Research opportunity

This review has illuminated the intricacies of human-AI interaction, particularly through a socio-emotional lens, underscoring the significance of trust, empathy, and rapport in augmenting human productivity. However, the research horizon in this domain remains vast and underexplored. Future studies should delve into the nuanced mechanisms of how socio-emotional attributes of AI influence various user demographics, considering cultural, age-related, and professional differences. There is a compelling opportunity to investigate the differential impacts of these interactions across diverse sectors such as healthcare, education, and customer service, where AI's role is rapidly expanding. Further, empirical research is needed to evaluate the long-term effects of sustained human-AI interactions on human psychological well-being and social behavior. This includes examining potential dependencies or over-reliance on AI for emotional support as well as exploring the ethical dimensions of human-AI relationships, especially in contexts where AI begins to substitute traditional human roles, warrants deeper inquiry.

There is also a pressing need for interdisciplinary research that connect insights from psychology, sociology, and AI technology to design AI systems that are not only technically proficient but also emotionally intelligent and culturally aware. Such research could pave the way for AI systems that are better aligned with human emotional and cognitive needs, thus enhancing their acceptance and effectiveness in collaborative settings. As importantly, AI continues to evolve and therefore investigating the potential for AI systems to not just mimic human emotions but to understand and appropriately respond to them in real-time scenarios presents an exciting frontier. This could significantly advance the development of AI as true socio-emotional partners in human interactions, leading to breakthroughs in personalized AI experiences and more profound human-AI collaborations. That will not only contribute to the academic discourse but also guide practical implementations, shaping a future where AI is an integral, empathetic, and responsive partner in various aspects of human life.

7 Conclusion

This research provides a comprehensive analysis of human interaction with artificial intelligence (AI), highlighting the critical role of socio-emotional attributes in enhancing human-AI collaboration. Using a GTLR methodology, we identified five key themes—rapport, trust, user engagement, empathy, and anthropomorphization—that are essential for aligning AI systems more closely with human emotional and cognitive needs, thereby improving collaborative efficiency and productivity.

Establishing a harmonious relationship based on mutual understanding and empathetic engagement is crucial (Cialdini and James, 2009). AI systems designed to recognize and respond to socio-emotional cues can significantly enhance user satisfaction and cooperation. Trust, encompassing both cognitive and emotional dimensions, reflects the user's confidence in the AI system's reliability, integrity, and competence. High levels of trust reduce perceived risks and increase user commitment. Emotional investment and sustained engagement with AI are influenced by the perceived usefulness and satisfaction derived from interactions. Effective AI design that meets user expectations fosters deeper commitment. AI systems capable of recognizing and appropriately responding to user emotions can foster a sense of understanding and connection, which is crucial for effective human-AI interactions. Attributing human characteristics to AI systems makes them more relatable and engaging, enhancing user acceptance and satisfaction.

The study underscores the necessity of a paradigm shift in AI development, moving from a primary focus on technical proficiency to a holistic approach that incorporates socio-emotional intelligence. This shift is essential for creating AI systems that are not only technically advanced but also capable of forming meaningful and productive collaborations with humans. The findings advocate for AI designs that prioritize emotional intelligence, leading to more effective and human-centric technological advancements. Such insights from this study are highly relevant across various sectors, including healthcare, education, and customer service. In healthcare, empathetic AI systems can improve patient trust and engagement, leading to better health outcomes. In education, AI tutors that build rapport with students can enhance learning experiences. In customer service, anthropomorphized AI can increase customer satisfaction and loyalty.

While this study offers significant contributions, it also highlights areas for future research. Further studies should explore the long-term effects of human-AI interactions on psychological well-being and social behavior. Additionally, there is a need for interdisciplinary research that bridges insights from psychology, sociology, and AI technology to design systems that are emotionally intelligent and culturally aware. There is also potential to extend the research to include additional keywords, such as the full term "artificial intelligence," which were initially deemed less relevant during the construction of the search strategy. Future research should consider including these and other terms to explore interdisciplinary connections more comprehensively, potentially expanding the search to include additional databases like Web of Science.

Incorporating socio-emotional attributes into AI design is pivotal for fostering productive and meaningful human-AI interactions. By prioritizing elements such as trust, empathy, rapport, user engagement, and anthropomorphization, AI systems can be more effectively integrated into human activities, leading to advancements that are both technologically innovative and human-centric. This research underscores the importance of continuous exploration and dialog in this domain, ensuring that AI advancements align with human dignity and societal welfare. The study's findings advocate for a comprehensive approach in AI development, one that equally values technological prowess and socio-emotional intelligence, to achieve a harmonious integration of AI into various facets of human life.

Author contributions

MK: Writing – original draft, Writing – review & editing. VP: Writing – original draft, Writing – review & editing. MS: Writing – original draft, Writing – review & editing. VJ: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ädel, A. (2011). Rapport building in student group work. *J. Pragmat.* 43, 2932–2947. doi: 10.1016/j.pragma.2011.05.007

Ammanabrolu, P., and Riedl, M. O. (2021). Situated language learning via interactive narratives. *Patterns* 2:100316. doi: 10.1016/j.patter.2021.100316

Araujo, T. (2018). Living up to the chatbot hype: the influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Comput. Hum. Behav.* 85, 183–189. doi: 10.1016/j.chb.2018.03.051

Argyle, M. (1990). The biological basis of rapport. *Psychol. Inq.* 1, 297–300. doi: 10.1207/s15327965pli0104_3

Attfield, S., Kazai, G., Lalmas, M., and Piwowarski, B. (2011). Towards a science of user engagement (position paper). In proceedings of the WSDM workshop on user modeling for web applications. 9–12. Available at: https://www.zdnet.com/article/alice-victorious-in-ai-challenge/

Avula, S., Chadwick, G., Arguello, J., and Capra, R. (2018). Searchbots: User engagement with chatbots during collaborative search. Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (pp. 52–61). ACM, New York.

Bellamy, E., and Beaumont, M. (2009). Looking backward, 2000–1887. Oxford: Oxford University Press.

Bickmore, T., and Picard, R. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transact. Comp. Human Inter.* 12, 293–327. doi: 10.1145/1067860.1067867

Bostrom, N., and Yudkowsky, E. (2014). "The ethics of artificial intelligence" in The Cambridge handbook of artificial intelligence. eds. K. Frankish and W. Ramsey (Cambridge, United Kingdom: Cambridge University Press), 316–334.

Boyle, E. A., Connolly, T. M., Hainey, T., and Boyle, J. M. (2012). Engagement in digital entertainment games: a systematic review. *Comput. Hum. Behav.* 28, 771–780. doi: 10.1016/j.chb.2011.11.020

Brown, E., and Cairns, P. (2004). "A grounded investigation of game immersion" in CHI'04 extended abstracts on human factors in computing systems (New York: ACM), 1297–1300.

Brynjolfsson, E., Horton, J. J., and Ozimek, A. (2020). COVID-19 and remote work: An early look at US data: National Bureau of Economic Research (NBER) - Massachusetts, USA.

Brynjolfsson, E., and McAfee, A. (2014). The book. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York, USA: W.W. Norton & Company.

Brynjolfsson, E., Rock, D., and Syverson, C. (2018). Artificial intelligence and the modern productivity paradox: a clash of expectations and statistics. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press. 23–57.

Buber, M. (2002) in Between man and man. ed. R. Gregor-Smith. *2nd* ed (Routledge).

Buber, M., and Smith, R. G. (1958). I and thou. Edinburgh: Clark.

Cervone, A., Gambi, E., Tortoreto, G., Stepanov, E. A., and Riccardi, G. (2018). Automatically predicting user ratings for conversational systems. Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it. 10, 12. Accademia University Press, Torino.

Cialdini, R. B., and James, L. (2009). Influence: Science and practice, vol. *4*. Boston: Pearson education.

Ciechanowski, L., Przegalinska, A., and Wegner, K. (2017). The necessity of new paradigms in measuring human-chatbot interaction. Proceedings of the International Conference on Applied Human Factors and Ergonomics. 205–214. Springer, Cham.

Clifford, C. (2016). Elon musk: Robots will take your jobs, government will have to pay your wage. CNBC. Available at: https://www.cnbc.com/2016/11/04/elon-musk-robots-will-take-your-jobs-government-will-have-to-pay-your-wage.html (Accessed 11. Oct. 2022).

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics Inf. Technol.* 12, 235–241. doi: 10.1007/s10676-010-9221-y

Davenport, T., Guha, A., and Grewal, D. (2020). How artificial intelligence will change the future of marketing. *J. Acad. Mark. Sci.* 48, 24–42. doi: 10.1007/s11747-019-00696-0

Davenport, T., and Kirby, J. (2016). Only humans need apply: Winners and losers in the age of smart machines. New York, USA: Harper Business.

de Graaf, M. M., Allouch, S. B., and Klamer, T. (2015). Sharing a life with Harvey: exploring the acceptance of and relationship-building with a social robot. *Comput. Hum. Behav.* 43, 1–14. doi: 10.1016/j.chb.2014.10.030

Debeauvais, T. (2016). Challenge and retention in games (Ph.D. dissertation, University of California, Irvine). ProQuest dissertations and theses. Available at: http://search.proquest.com.libaccess.sjlibrary.org/docview/1808939056?accountid=10361 (Accessed January 25, 2024).

Driskell, T., Blickensderfer, E. L., and Salas, E. (2013). Is three a crowd? Examining rapport in investigative interviews. *Group Dyn. Theory Res. Pract.* 17, 1–13. doi: 10.1037/a0029686

Fadhil, A., Schiavo, G., Wang, Y., and Yilma, B. A. (2018). The effect of emojis when interacting with a conversational interface assisted health coaching system. Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare (pp. 378–383). ACM, New York.

Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Mental Health* 4:e19. doi: 10.2196/mental.7785

Fleming, K., Sundararajan, A., Dhar, V., Siebel, T., Slaughter, A., Wald, J., et al. (2020). The digital future of work: What skills will be needed? McKinsey Global Institute.

Floridi, L., and Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Sci. Review* 1, 1–15. doi: 10.1162/99608f92.8cd550d1

Floridi L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Mind. Mach.* 28, 689–707. doi: 10.1007/s11023-018-9482-5

Følstad, A., Nordheim, C. B., and Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? An exploratory interview study. Proceedings of the International Conference on Internet Science (pp. 194–208). Springer, Cham.

Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robot. Auton. Syst.* 42, 143–166. doi: 10.1016/S0921-8890(02)00372-X

Ford, M. (2015). Rise of the robots: Technology and the threat of a jobless future. New York, USA: Basic Books.

Fortunati, L., and Edwards, A. (2020). Opening space for theoretical, methodological, and empirical issues in human-machine communication. *Hum. Mach. Commun.* 1, 7–18. doi: 10.30658/hmc.1.1

Frey, C. B., and Osborne, M. A. (2017). The future of employment: how susceptible are jobs to computerization? *Technol. Forecast. Soc. Chang.* 114, 254–280. doi: 10.1016/j.techfore.2016.08.019

Frison, A. K., Wintersberger, P., Riener, A., Schartmüller, C., Boyle, L. N., Miller, E., et al. (2019). In UX we trust: investigation of aesthetics and usability of driver-vehicle interfaces and their impact on the perception of automated driving. Proceedings of the 2019 CHI conference on human factors in computing systems CHI 2019, Glasgow, Scotland, UK. New York: ACM. 1–13.

Garnham, A. (1987). Artificial intelligence: An introduction. London, United Kingdom: Routledge.

Gasson, S. (2009). "Employing a grounded theory approach for MIS research" in Handbook of research on contemporary theoretical models in information systems. eds. Y. K. Dwivedi, B. Lal, M. D. Williams, S. L. Schneberger and M. Wade (Hershey, Pennsylvania, USA: IGI Global), 34–56.

Gilbert, D. T. (1991). How mental systems believe. *Am. Psychol.* 46, 107–119. doi: 10.1037/0003-066X.46.2.107

Glaser, B. G., and Strauss, A. L. (2017). Discovery of grounded theory: Strategies for qualitative research. London, United Kingdom: Routledge.

Goethe, O., Salehzadeh, Niksirat K., Hirskyj-Douglas, I., Sun, H., Law, E.L.C., and Ren, X. (2019). From UX to engagement: Connecting theory and practice, addressing ethics and diversity. Proceedings of the International Conference on Human-Computer Interaction (pp. 91–99). Springer, Cham.

Goodwins, R. (2001). ALICE victorious in AI challenge. ZDNet. Available at: https://www.zdnet.com/article/alice-victorious-in-ai-challenge/ (Accessed February 6, 2023).

Gremler, D., and Gwinner, K. (2008). Rapport-building behaviors used by retail employees. *J. Retail.* 84, 308–324. doi: 10.1016/j.jretai.2008.07.001

Grudin, J., and Jacques, R. (2019). Chatbots, humbots, and the quest for artificial general intelligence. Proceedings of the 2019 CHI conference on human factors in computing systems. New York, USA: ACM Press. 1–11.

Guzman, A. L. (2018). "What is human-machine communication, anyway?" in Human machine communication: rethinking communication, technology, and ourselves. ed. A. L. Guzman, New York, USA: Peter Lang. 1–28.

Guzman, A. L. (2020). Ontological boundaries between humans and computers and the implications for human-machine communication. *Hum. Mach. Commun.* 1, 37–54. doi: 10.30658/hmc.1.3

Ho, C. C., and MacDorman, K. F. (2017). Measuring the uncanny valley effect. *Int. J. Soc. Robot.* 9, 129–139. doi: 10.1007/s12369-016-0380-9

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat. Mach. Intellig.* 1, 389–399. doi: 10.1038/s42256-019-0088-2

Kahneman, D. (2011). Thinking, fast and slow. Straus and Giroux: Farrar.

Kaplan, J. (2015). Humans need not apply: A guide to wealth and work in the age of artificial intelligence. New Haven, Connecticut, USA: Yale University Press.

Knafo, D. (2015). Guys and dolls: relational life in the technological era. *Psychoanal. Dialog.* 25, 481–502. doi: 10.1080/10481885.2015.1055174

Lee, K. M. (2004). Why presence occurs: evolutionary psychology, media equation, and presence. *Presence* 13, 494–505. doi: 10.1162/1054746041944830

Littell, J. H., Corcoran, J., and Pillai, V. (2008). Systematic reviews and meta-analysis. New York, USA: Oxford University Press.

Liu, D., Santhanam, R., and Webster, J. (2017). Toward meaningful engagement: a framework for design and research of gamified information systems. *MIS Q.* 41, 1011–1034. doi: 10.25300/MISQ/2017/41.4.01

Liu, B., and Sundar, S. S. (2018). Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychol. Behav. Soc. Netw.* 21, 625–636. doi: 10.1089/cyber.2018.0110

Lombard, M., and Ditton, T. (1997). At the heart of it all: the concept of presence. *J. Comput.-Mediat. Commun.* 3:JCMC321. doi: 10.1111/j.1083-6101.1997.tb00072.x

Lombard, M., and Xu, K. (2021). Social responses to media technologies in the 21st century: the media are social actors paradigm. *Hum. Mach. Commun.* 2, 29–55. doi: 10.30658/hmc.2.2

Lukoff, K., Yu, C., Kientz, J., and Hiniker, A. (2018). What makes smartphone use meaningful or meaningless? *Proceed. ACM Interact. Mobile Wearable Ubiquitous Technol.* 2, 1–26. doi: 10.1145/3191754

MacInnis, D., and Folkes, V. S. (2017). Humanizing brands: when brands seem to be like me, part of me, and in a relationship with me. *J. Consum. Psychol.* 27, 355–374. doi: 10.1016/j.jcps.2016.12.003

Manyika, J., Chui, M., Miremadi, M., Bughin, J., George, K., Willmott, P., et al. (2017). A future that works: Automation, employment, and productivity: McKinsey Global Institute. Available at: https://www.mckinsey.com/featured-insights/digital-disruption/harnessing-automation-for-a-future-that-works/de-DE (Accessed February 2, 2024).

Mattarelli, E., Bertolotti, F., and Macrì, D. M. (2013). The use of ethnography and grounded theory in the development of a management information system. *Eur. J. Inf. Syst.* 22, 26–44. doi: 10.1057/ejis.2011.34

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manag. Rev.* 20, 709–734. doi: 10.2307/258792

McKinsey Global Institute. (2020). The future of work in Europe: automation, workforce transitions, and the shifting geography of employment. McKinsey & Company. Available at: https://www.mckinsey.com/featured-insights/future-of-work/the-future-of-work-in-europe (Accessed February 2, 2024)

Mehrabian, A., and Epstein, N. (1972). A measure of emotional empathy. *J. Pers.* 40, 525–543. doi: 10.1111/j.1467-6494.1972.tb00078.x

Mencarini, E., Rapp, A., Tirabeni, L., and Zanacanaro, M. (2019). Designing wearable Systems for Sport: a review of trends and opportunities in human-computer interaction. *IEEE Transact. Hum. Mach. Syst.* 49, 314–325. doi: 10.1109/THMS.2019.2919702

Menzel, P., and D'Aluisio, F. (2000). Robo sapiens: Evolution of a new species. Cambridge, MA: MIT Press.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intell. Syst.* 21, 18–21. doi: 10.1109/MIS.2006.80

Morrison, R. L. (2009). Are women tending and befriending in the workplace? Gender differences in the relationship between workplace friendships and organizational outcomes. *Sex Roles* 60, 1–13. doi: 10.1007/s11199-008-9513-4

Morrissey, K., and Kirakowski, J. (2013). 'Realness' in chatbots: Establishing quantifiable criteria. Proceedings of the International Conference on Human-Computer Interaction (pp. 87–96). Springer, Berlin, Heidelberg.

Nass, C., and Moon, Y. (2000). Machines and mindlessness: social responses to computers. *J. Soc. Issues* 56, 81–103. doi: 10.1111/0022-4537.00153

Niculescu, A., van Dijk, B., Nijholt, A., Li, H., and See, S. L. (2013). Making social robots more attractive: the effects of voice pitch, humor and empathy. *Int. J. Soc. Robot.* 5, 171–191. doi: 10.1007/s12369-012-0171-x

Nunes, F., Verdezoto, N., Fitzpatrick, G., Kyng, M., Grönvall, E., and Storni, C. (2015). Self-care technologies in HCI: trends, tensions, and opportunities. *ACM Transact. Comp. Hum. Interact.* 22, 1–45. doi: 10.1145/2803173

O'Brien, H., and Toms, E. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *J. Amer. Soc Inform. Sci. Technol.* 59, 938–955. doi: 10.1002/asi.20801

Peters, C., Castellano, G., and de Freitas, S. (2009). "An exploration of user engagement in HCI" in Proceedings of the international workshop on affective-aware virtual agents and social robots (New York: ACM), 1–3. doi: 10.1145/1655260.1655269

Portela, M., and Granell-Canut, C. (2017). A new friend in our smartphone? Observing interactions with Chatbots in the search of emotional engagement. Proceedings of the XVIII international conference on human computer interaction. New York, USA: ACM Press. 7, 1–7.

Ren, X. (2016). Rethinking the relationship between humans and computers. *IEEE Comp.* 49, 104–108. doi: 10.1109/MC.2016.253

Ruan, S., Jiang, L., Xu, J., Tham, B. J. K., Qiu, Z., Zhu, Y., et al. (2019). Quizbot: A dialogue-based adaptive learning system for factual knowledge. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–13). ACM, New York.

Sabeti, S., Shoghli, O., Baharani, M., and Tabkhi, H. (2021). Toward AI-enabled augmented reality to enhance the safety of highway work zones: feasibility, requirements, and challenges. *Adv. Eng. Inform.* 50:101429. doi: 10.1016/j.aei.2021.101429

Saket, B., Endert, A., and Stasko, J. (2016). Beyond usability and performance: A review of user experience-focused evaluations in visualization. Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization - BELIV 16 (pp. 133–142). ACM, New York.

Schlaepfer, R., and Koch, M. (2015) Industry 4.0: Challenges and solutions for the digital transformation and use of exponential technologies. Deloitte. Available at: https://www2.deloitte.com/content/dam/Deloitte/ch/Documents/manufacturing/ch-en-manufacturing-industry-4-0-24102014.pdf (Accessed 6 July 2022).

Schwab, K. (2016). The fourth industrial revolution. New York, USA: Crown Business.

Schwind, V., Wolf, K., and Henze, N. (2018). Avoiding the uncanny valley in virtual character design. *Interactions* 25, 45–49. doi: 10.1145/3236673

Sheehan, B., Jin, H. S., and Gottlieb, U. (2020). Customer service chatbots: anthropomorphism and adoption. *J. Bus. Res.* 115, 14–24. doi: 10.1016/j.jbusres.2020.04.030

Short, E., Hart, J., Vu, M., and Scassellati, B. (2010). "No fair!! An interaction with a cheating robot" in In 2010 5th ACM/IEEE international conference on human-robot interaction (HRI) (IEEE), New Haven, Connecticut, USA: Yale University. 219–226.

Siau, K., and Wang, W. (2018). Building Trust in Artificial Intelligence, machine learning, and robotics. *Cutter Bus. Technol. J.* 31, 47–53.

Skjuve, M., Haugstveit, I. M., Følstad, A., and Brandtzaeg, P. B. (2019). Help! Is my chatbot falling into the uncanny valley? An empirical study of user experience in human-chatbot interaction. *Hum. Technol.* 15, 30–54. doi: 10.17011/HT/URN.201902201607

Smith, A. (2014). AI, robotics, and the future of jobs. [Online]. Pew Research Center: Internet, Science & Tech. Available at: http://www.pewinternet.org/2014/08/06/future-of-jobs (Accessed November 7, 2022).

Stein, J. P., and Ohler, P. (2017). Venturing into the uncanny valley of mind—the influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition* 160, 43–50. doi: 10.1016/j.cognition.2016.12.010

Stern, A. (2016). Raising the floor: How a universal basic income can renew our economy and rebuild the American dream. New York, USA: PublicAffairs.

Stronks, B., Nijholt, A., van der Vet, P. E., Heylen, D., and Limburg, D. O. (2002). Friendship relations with embodied conversational agents: Integrating social psychology in ECA design Report. Enschede, The Netherlands: Parlevink Research Group, University of Twente.

Sundar, S. S., Bellur, S., Oh, J., Jia, H., and Kim, H. S. (2016). Theoretical importance of contingency in human-computer interaction: effects of message interactivity on user engagement. *Commun. Res.* 43, 595–625. doi: 10.1177/0093650214534962

Svenningsson, N., and Faraon, M. (2019). Artificial intelligence in conversational agents: A study of factors related to perceived humanness in Chatbots. Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference (pp. 151–161). ACM, New York.

Syam, N., and Sharma, A. (2018). Waiting for a sales renaissance in the fourth industrial revolution: machine learning and artificial intelligence in sales research and practice. *Ind. Mark. Manag.* 69, 135–146. doi: 10.1016/j.indmarman.2017.12.019

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., et al. (2020). User experiences of social support from companion Chatbots in everyday contexts: thematic analysis. *J. Med. Internet Res.* 22:e16235. doi: 10.2196/16235

ter Stal, S., Kramer, L. L., Tabak, M., op den Akker, H., and Hermens, H. (2020). Design features of embodied conversational agents in eHealth: a literature review. *Int. J. Hum.-Comput. Stud.* 138:102409. doi: 10.1016/j.ijhcs.2020.102409

The Register. (2024). Deepfake CFO tricks Hong Kong biz out of $25 million. Available at: https://www.theregister.com/2024/02/04/deepfake_cfo_scam_hong_kong/ (Accessed January 12, 2024).

Turkle, S. (2011). Alone together: Why we expect more from technology and less from each other. New York, USA: Basic Books.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases: biases in judgments reveal some heuristics of thinking under uncertainty. *Science* 185, 1124–1131. doi: 10.1126/science.185.4157.1124

Urakami, J., Moore, B. A., Sutthithatip, S., and Park, S. (2019). Users' perception of empathic expressions by an advanced intelligent system. Proceedings of the 7th International Conference on Human-Agent Interaction (pp. 11–18). ACM, New York.

Vossen, H. G., Piotrowski, J. T., and Valkenburg, P. M. (2015). Development of the adolescent measure of empathy and sympathy (AMES). *Personal. Individ. Differ.* 74, 66–71. doi: 10.1016/j.paid.2014.09.040

West, D. M. (2018). The future of work: Robots, AI, and automation. Washington, D.C., USA: Brookings Institution Press.

Westerman, D., Cross, A. C., and Lindmark, P. G. (2019). I believe in a thing called bot: perceptions of the humanness of "chatbots". *Commun. Stud.* 70, 295–312. doi: 10.1080/10510974.2018.1557233

Wolfswinkel, J. F., Furtmueller, E., and Wilderom, C. P. (2013). Using grounded theory as a method for rigorously reviewing literature. *Eur. J. Inf. Syst.* 22, 45–55. doi: 10.1057/ejis.2011.51

Xiao, L., and Kumar, V. (2021). Robotics for customer service: a useful complement or an ultimate substitute? *J. Serv. Res.* 24, 9–29. doi: 10.1177/1094670519878881

Xiao, Z., Zhou, M. X., Chen, W., Yang, H., and Chi, C. (2020). If I hear you correctly: Building and evaluating interview Chatbots with active listening skills. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–14). ACM, New York.

Yen, C., and Chiang, M. C. (2020). Trust me, if you can: a study on the factors that influence consumers' purchase intention triggered by chatbots based on brain image evidence and self-reported assessments. *Behav. Inform. Technol.* 40, 1177–1194. doi: 10.1080/0144929X.2020.1743362

Yoganathan, V., Osburg, V.-S., and Kunz, W. H. (2021). Check-in at the Robo desk: effects of automated social presence on social cognition and service implications. *Tour. Manag.* 85:104309. doi: 10.1016/j.tourman.2021.104309

Zamora, J. (2017). I'm sorry, Dave, I'm afraid I can't do that: Chatbot perception and expectations. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. New York, USA. 253–260.

Zhao, S. (2006). Humanoid social robots as a medium of communication. *New Media Soc.* 8, 401–419. doi: 10.1177/1461444806061951

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 17 October 2024
DOI 10.3389/frai.2024.1397860

Check for updates

OPEN ACCESS

EDITED BY
Simone Belli,
Complutense University of Madrid, Spain

REVIEWED BY
Raymond Lee,
Beijing Normal University-Hong Kong Baptist University United International College, China
Christian Lebiere,
Carnegie Mellon University, United States

*CORRESPONDENCE
Kazuma Nagashima
✉ nagashima.kazuma.16@shizuoka.ac.jp
Junya Morita
✉ j-morita@inf.shizuoka.ac.jp

RECEIVED 08 March 2024
ACCEPTED 30 September 2024
PUBLISHED 17 October 2024

CITATION
Nagashima K, Morita J and Takeuchi Y (2024) Intrinsic motivation in cognitive architecture: intellectual curiosity originated from pattern discovery. *Front. Artif. Intell.* 7:1397860.
doi: 10.3389/frai.2024.1397860

Intrinsic motivation in cognitive architecture: intellectual curiosity originated from pattern discovery

Kazuma Nagashima[1]*, Junya Morita[1,2]* and Yugo Takeuchi[1]

[1]Department of Information Science and Technology, Graduate School of Science and Technology, Shizuoka University, Hamamatsu, Japan, [2]Department of Behavior Informatics, Faculty of Informatics, Shizuoka University, Hamamatsu, Japan

Studies on reinforcement learning have developed the representation of curiosity, which is a type of intrinsic motivation that leads to high performance in a certain type of tasks. However, these studies have not thoroughly examined the internal cognitive mechanisms leading to this performance. In contrast to this previous framework, we propose a mechanism of intrinsic motivation focused on pattern discovery from the perspective of human cognition. This study deals with intellectual curiosity as a type of intrinsic motivation, which finds novel compressible patterns in the data. We represented the process of continuation and boredom of tasks driven by intellectual curiosity using "pattern matching," "utility," and "production compilation," which are general functions of the adaptive control of thought-rational (ACT-R) architecture. We implemented three ACT-R models with different levels of thinking to navigate multiple mazes of different sizes in simulations, manipulating the intensity of intellectual curiosity. The results indicate that intellectual curiosity negatively affects task completion rates in models with lower levels of thinking, while positively impacting models with higher levels of thinking. In addition, comparisons with a model developed by a conventional framework of reinforcement learning (intrinsic curiosity module: ICM) indicate the advantage of representing the agent's intention toward a goal in the proposed mechanism. In summary, the reported models, developed using functions linked to a general cognitive architecture, can contribute to our understanding of intrinsic motivation within the broader context of human innovation driven by pattern discovery.

KEYWORDS

cognitive modeling, ACT-R, intrinsic motivation, intellectual curiosity, pattern discovery

1 Introduction

According to Baron-Cohen (2020), human evolution and the development of civilization are associated with "systematizing mechanisms," which are achieved by discovering, combining, and using patterns of cause-and-effect relationships in an environment. He also stated that the ability of humans to think systematically has evolved by using the "if-and-then" logic to combine patterns, resulting in inventions and innovations that lead to our modern society.

Several studies have reported that such an ability of pattern discovery is associated with fun, a personal feeling leading to intrinsic motivation (Caillois, 1958; Csikszentmihalyi, 1990; Huizinga, 1939; Koster, 2013). The other researchers (Aubret et al., 2019; Schmidhuber, 2010) have also explored the computational realization of intrinsic motivation employing the framework of reinforcement learning (Sutton and Barto, 1998). However, these studies have not explored the link between intrinsic motivation and primitive cognitive functions related to pattern discovery. Therefore, further analysis of the computational mechanisms of intrinsic motivation in terms of agents' internal processing is needed.

The aforementioned problem can be addressed by using a cognitive architecture that integrates the basic cognitive functions involved in various tasks. Despite the existence of several cognitive architectures, the architectural differences have been reduced over the years and integrated into a common structure (Laird et al., 2017). The representative architecture adopting such a structure is adaptive control of thought-rational (ACT-R), developed by Anderson (2007). According to Kotseruba and Tsotsos (2020)'s comprehensive review of the topic, ACT-R is one of the most widely used cognitive architectures, including a greater number of features compared to the other architectures.

In this study, we propose a mechanism of intrinsic motivation based on pattern discovery by integrating primitive cognitive functions of ACT-R. The main advantage of the proposed approach is its interpretability. Based on commonly used building blocks in the architecture, our proposed mechanism can provide a foundation for understanding intrinsic motivation from the perspective of human cognition. Furthermore, this study presents a simulation experiment to explore the conditions of stimulating intrinsic motivation and the learning process driven by stimulated intellectual curiosity. Our analysis confirmed that the proposed mechanism can represent the role of intellectual motivation in human learning at diverse levels of thinking and task difficulty. Additionally, we examined the relationship between the proposed mechanism and an existing mechanism of intrinsic motivation based on reinforcement learning.

The remainder of this paper is organized as follows. Section 2 summarizes the existing studies related to this concept. Section 3 introduces the proposed mechanism, which is developed based on pattern discovery. The effectiveness of the mechanism is discussed based on simulations in Section 4. Finally, Section 5 summarizes the findings and indicates directions for future investigations.

2 Related works

The objective of this study is to represent a mechanism of intrinsic motivation based on the discovery of patterns. This section focuses on three directions of previous studies, namely, human curiosity, machine curiosity, and cognitive models with cognitive architectures.

2.1 Human curiosity

Numerous studies have attempted to systematize intrinsic motivation as a driving factor to continue activities in a wide range of fields, including education, entertainment, healthcare, sports, and work. For instance, Malone (1981), who tried to systematize this concept in entertainment fields, categorized intrinsic motivation into three types, namely, "challenge," which originates from goals of appropriate difficulty; "fantasy," which leads to the imagination of unrealistic experiences; and "curiosity," which is stimulated by a surprising, interesting, or fun activity. Here, curiosity is related to the discussion presented in Section 1 that pattern discovery accompanying the feeling of fun has led to human innovations. However, we believe that the first type of intrinsic motivation, challenge, is inseparable from curiosity. Rather than treating those as independent factors, we assume that curiosity is a mechanism of intrinsic motivation, stimulated by the appropriate difficulty (challenge) of a task.

The above assumption is supported by several authors who reported the relationship between the levels of task difficulty, the preferred level of thinking, and intrinsic motivation. The theory behind this is referred to as the optimal level of intrinsic motivation (Csikszentmihalyi, 1990; Yerkes and Dodson, 1908). According to this theory, intrinsic motivation is effectively stimulated when the task difficulty level matches the preferred level of thinking of a person. Furthermore, the level of thinking can be located on an axis with at least two levels. These include a shallow automatic level without careful thinking (fast process) and a deep deliberative level that requires time to carefully think (slow process) (Brooks, 1986; Evans, 2003; Kahneman, 2011).

2.2 Machine curiosity

Based on the aforementioned discussion, we assumed a close relationship between curiosity and the feeling of fun involved in the discovery of patterns. This relationship was computationally theorized by Schmidhuber (2010), wherein the discovery of patterns is defined as identifying and compressing recurring canonical patterns in data. Schmidhuber also related compressing data or obtaining compressible data to fun by assuming that the agent aims to maximize fun as a reward. This idea was based on the prediction error theory (Friston, 2010), which considers curiosity to be caused by the difference between prior predictions and the current situation (Bayesian surprise). In Schmidhuber's theory, prediction implies applying already compressed data; here, surprise occurs when identifying a pattern that can be newly compressed.

Schmidhuber's proposal can be discussed as an extension of conventional reinforcement learning. Typically, agents in reinforcement learning receive rewards from the environment and intend to maximize them over time. Sutton and Barto (1998) distinguished the boundaries between the agent and the environment from the physical boundaries between the body and the environment. Based on this idea, Singh et al. (2005) proposed a framework of intrinsically motivated reinforcement learning (IMRL), which divides the environment into external and internal segments. In contrast to conventional reinforcement learning, which directly receives a reward from the external environment, rewards in IMRL are determined depending on the state of the internal environment, such as stimulating curiosity for an unexpected response.

Since the proposal of IMRL, the framework of reinforcement learning has significantly progressed by integrating deep learning techniques. The preliminary framework was referred to as Deep Q Network (DQN) (Mnih et al., 2015), which combined Q learning with a convolutional neural network (CNN). Subsequently, several researchers introduced the concept of intrinsic motivation in deep reinforcement learning. For instance, Bellemare et al.

(2016) developed count-based exploration methods, wherein visit counts were used to guide an agent's behavior toward reducing uncertainty. In their research, calculating prediction errors as internal rewards led the agents to search for novel states and ultimately outperformed DQN. Following this idea, Pathak et al. (2017) proposed the intrinsic curiosity module (ICM), which regarded the difference between the predicted state of an agent and the situation obtained from the pixel information on the screen as curiosity. Herein, ICM was integrated with the asynchronous actor-critic model (A3C) (Mnih et al., 2016). Based on this method, Burda et al. (2018) implemented an approach to explore the environment using only internal rewards regarding curiosity. Moreover, Burda et al. (2019) proposed a method named random network distillation, which made it possible to learn tasks that were difficult to accomplish with the previous methods.

2.3 Cognitive models with cognitive architectures

Although the aforementioned studies successfully represented curiosity in reinforcement learning, their integration with cognitive functions has not been sufficiently explored. As explained in Section 2.1, curiosity is associated with the discovery of patterns. Therefore, the computational representation should include basic human cognitive functions behind pattern discovery; this can be achieved using ACT-R. The subsequent section explains the representation of individual cognitive functions in ACT-R and the type of learning realized by combining cognitive functions. Herein, we predominantly focus on the cognitive functions of ACT-R involved in this study. Further information on ACT-R can be obtained from Anderson (2007), the ACT-R manual (Bothell, 2020), and other reviews (Ritter et al., 2019).

2.3.1 Structure of ACT-R modules

ACT-R comprises modules corresponding to brain regions, as indicated in Figure 1. The mapping between the modules and regions has been discussed based on neuroscientific findings (Stocco et al., 2021). The principal assumption of this structure is that one module takes responsibility for a set of functions. For instance, the declarative module comprises functions for storing experience and knowledge, the imaginal module contains functions to create new knowledge by combining multiple internal representations, and the function of the goal module is to maintain the current status of tasks to manage the process of the model. The state of each module at each time point (e.g., the declarative knowledge being recalled and the state of the current goal) is expressed using a symbol referred to as a *chunk*, which is stored in a buffer for each module. The chunks stored in the buffer are evaluated using a type of procedural knowledge, called "productions," comprising IF (conditions) and THEN (actions) clauses in the production module. The productions transmit chunks describing commands to modules as actions, such as searching for knowledge that satisfies the conditions and updating the current state of the task.

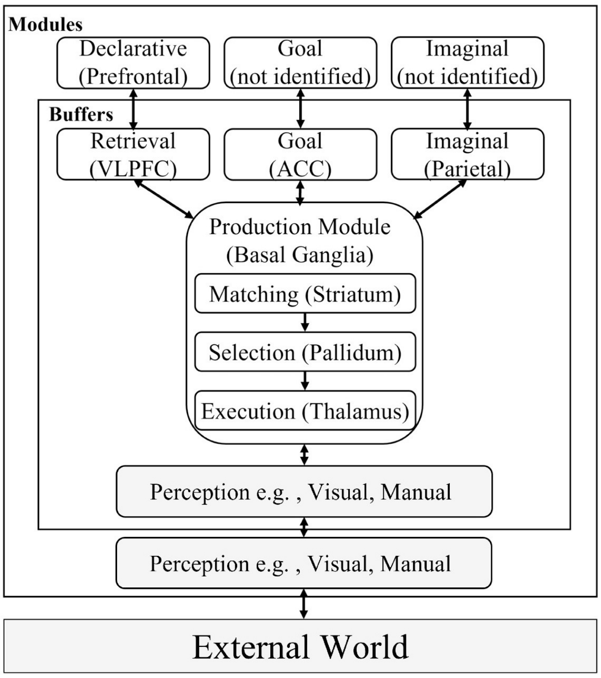


FIGURE 1
Overview of the adaptive control of thought-rational (ACT-R) modules. Modules excluded in this study are grayed out. This figure is created with reference to Anderson et al. (2004) and Ritter et al. (2019). VLPFC, ventrolateral prefrontal cortex; ACC, anterior cingulate cortex; DLPFC, Dorsolateral prefrontal cortex; PPC, posterior parietal cortex.

Therefore, the declarative and production modules in ACT-R contain different types of knowledge. The retrieval cost of declarative knowledge (chunks) in the declarative module is greater than that of procedural knowledge (productions) in the production module. The cost in ACT-R corresponds to the processing time, which simulates human reaction times (van der Velde et al., 2022). A single production can be executed in 50 ms, whereas the retrieval of declarative knowledge requires longer as various factors are involved. Moreover, declarative knowledge is not automatically retrieved from the goal module or the external environment as it is always used by applying two productions; one for retrieving declarative knowledge and the other for applying the retrieved knowledge to change the states of buffers (e.g., goal or perceived external environment).

Biologically, the ACT-R theory assumes that the two types of knowledge are connected through the cortico-basal ganglia loop. As depicted in Figure 1, the productions are assumed to be executed in the basal ganglia; however, the ones used for retrieving declarative knowledge require the prefrontal cortex as well. Figure 2 illustrates a simple example of retrieving and using declarative knowledge through two productions. In the figure, variables "var1" and "var2" in the productions are bound to numerical values, such as 1 and 2, stored in the declarative module. This mechanism is referred to as "pattern matching" and is assumed to be executed intentionally in the prefrontal cortex [particularly in the ventrolateral prefrontal cortex (VLPFC) indicated in Figure 1]. Therefore, we considered

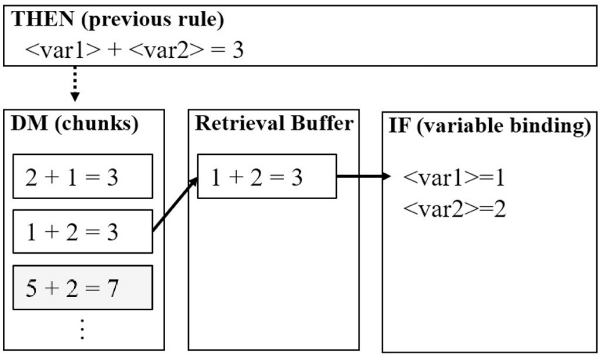


FIGURE 2
Simple example of pattern matching in the adaptive control of thought-rational (ACT-R) architecture. This example illustrates the flow of a declarative knowledge query to the declarative module (DM) in the THEN clause of the previous production. The variables are bound in the IF clause of the subsequent production.

the pattern matching between the current situation (buffers) and knowledge in the declarative module as the criterion for distinguishing the aforementioned levels of thinking (Section 2.1). In this framework, the shallow level of thinking involved fewer pattern-matching scenarios than the deliberative level of thinking.

2.3.2 Learning in ACT-R

The existence of pattern matching also makes a distinction between two types of learning in ACT-R: learning with pattern matching and learning without pattern matching. The latter type uses "utility learning," which corresponds to reinforcement learning. Specifically, it changes the selection probability of productions that conflict with each other by receiving rewards from the environment. Many studies have used this type of learning in ACT-R modeling (Anderson et al., 1993; Balaji et al., 2023; Ceballos et al., 2020; Xu and Stocco, 2021). For example, Fu and Anderson (2006) developed a model to solve the repeated maze task by applying procedural knowledge representing up-down and left-right movements. The model received positive rewards for actions that led to the achievement of the current goal and negative rewards for actions that failed to achieve the goal. As a result of their simulation, the model was able to learn optimal behavior in the maze search by repeating the rewarding trials.

The other type of learning in ACT-R involves pattern matching to retrieve chunks in the declarative module, which is called instance-based learning (IBL) (Gonzalez et al., 2003). This framework accumulates past problem-solving instances in the declarative module and uses it for future task trials. Several studies show that IBL outperforms conventional utility learning. Relating to this method, Reitter and Lebiere (2010) constructed a model to solve maze like Fu and Anderson (2006), but unlike them, by combining path-finding with backtracking and instance-based inference. In their model, location information of the maze was represented as declarative knowledge to construct a topological map (graph-like structure representing geological locations). In addition to conventional knowledge-search algorithms (e.g., depth-first search), an instance-based inference was applied by using stored maze-solving experience in the declarative module. By conducting simulations using the strategies of maze search, they demonstrated the advantage of this memory-based search.

Furthermore, ACT-R contains another learning function that uses the two aforementioned functions. This function is the "compilation," which combines two productions into a single production (Taatgen and Lee, 2003). During the task execution, this function integrates a repeatedly selected series of productions and reduces the number of productions used in the task as learning progresses. Typically, the target series of compilation involves pattern matching to retrieve declarative knowledge (Figure 2). The function replaces the variables present in the production with instantiated values in the declarative knowledge. Additionally, the conflicting conditions for pre-compiled and post-compiled productions are resolved using the utility learning. The post-compiled production inherits higher utility from those associated with the two pre-compiled productions. Furthermore, the utility of post-compiled production increases with the compilation of the same series of productions. This process increases the probability of selecting a post-compiled production to represent a routine and automatic operation (procedural knowledge) in a task.

2.3.3 Emotion in ACT-R

The subject of the present study, motivation, is considered part of the emotional or affective phenomena in the recently emerging field of affective science.[1] In this field, researchers have repeatedly pointed to the relations between emotions, cognition, and body (Barrett, 2017; Damasio, 2003; LeDoux and Pine, 2016), underscoring the importance of incorporating emotional and physiological responses into cognitive models.

Following these trends in affective science, several researchers have constructed ACT-R models that represent the interactions between cognition, emotion, and the body. For example, van Vugt and van der Velde (2018) constructed a model explaining depression based on the proportion of memories accompanied by emotional moods. Similarly, Juvina et al. (2018) considered the relationship between emotional memories and reward functions. In addition to these links between emotion and cognition, researchers have included psychophysiological factors such as fatigue (Atashfeshan and Razavi, 2017; Gunzelmann et al., 2009) and stress (Dancy et al., 2015) in ACT-R. Based on these models of emotions, several ACT-R models of motivations have been developed (Nishikawa et al., 2022; Nagashima et al., 2022; Yang and Stocco, 2024). Furthermore, in recent discussions on the common cognitive model, Rosenbloom et al. (2024) proposed an architecture including metacognitive modules to represent interactions between cognition and emotion.

However, to implement such emotional processes, all the aforementioned studies developed novel modules or functions of ACT-R. By contrast, the current study aims to model intrinsic motivation using the existing built-in functions of ACT-R. While

[1] The mission of the society for affective science includes "motivated states." See https://society-for-affective-science.org/about-sas/.

we recognize the importance of developing new modules to create a neurally faithful structure, we believe that, in line with the philosophy of cognitive architecture (Anderson et al., 2004), it is preferable to represent various cognitive processes by integrating a small set of core functions.

3 Mechanism of intellectual curiosity based on pattern discovery

This section proposes a mechanism of intrinsic motivation. Before presenting details of the mechanism, the basic idea behind our proposal is introduced.

3.1 Basic idea

The mechanism proposed here focuses on intellectual curiosity among the types of intrinsic motivation. We used the modifier "intellectual" based on the discussion reported by Malone (1981). According to him, the curiosity derived from higher-cognitive functions is distinguished from that derived from sensory perceptions. He further argued that the former initiates "a desire to bring better form to one's knowledge structures." This discussion is consistent with the principle of fun discussed by Schmidhuber (2010), who claimed that discovering the compressible structure of data would be beneficial to organize knowledge structures in the agent.

We developed a mechanism of intellectual curiosity by associating ACT-R pattern-matching computation. As explained earlier, pattern matching of ACT-R is a core mechanism for understanding higher-order cognitive processes with the discovery of structures (patterns) that map data (declarative knowledge) to a current situation (buffer states of the module) according to a pattern of variables in the production. This mechanism has been considered essential for achieving cognitive flexibility that adapts changing environments by leveraging existing knowledge in novel forms (Spiro et al., 2012). In fact, Anderson (2007) demonstrated that ACT-R could model human-specific cognitive functions, such as linguistic processing, metacognition, and analogical reasoning by using a certain type of pattern matching.[2] More importantly, ACT-R pattern matching is involved in the learning mechanisms, as described in Section 2. In the following part of this section, the learning mechanisms of ACT-R are combined into a general framework of intrinsic motivation.

3.2 Components of intellectual curiosity

To understand the role of intellectual curiosity in general cognitive processes, we first discuss its decay (boredom) process. Typically, boredom is caused by stimulus saturation and is related to learning processes as suggested by Csikszentmihalyi (1990). According to his theory, boredom occurs when the person extensively learns a particular task and it becomes less challenging. Raffaelli et al. (2018) reviewed research confirming such a process based on subjective and physiological indices, which sometimes showed complex interactions between cognitive and physiological processes. Based on these discussions especially about the relation between learning and boredom, we used the "utility learning" and "production compilation" to represent the decay of intellectual curiosity. Although the general concept of these mechanisms has already been discussed, the subsequent sections focus on the technical details of the modules as ingredients of our integrated mechanism of intellectual curiosity.

3.2.1 Motivation as utility for task continuation

We used utility learning in this study as a mechanism for determining whether a task should be continued or terminated. As mentioned in Section 2.3, the utility learning corresponds to reinforcement learning (Fu and Anderson, 2006). When multiple productions (i.e., the production for task continuation and the production for task termination) match the current situation, the probability of selecting a production can be calculated as

$$P(i) = \frac{e^{U_i/\sqrt{2s}}}{\sum_j e^{U_j/\sqrt{2s}}}, \tag{1}$$

where e denotes the base of the natural logarithm, s indicates the parameter that determines the variance of noise according to the logistic distribution, and j distinguishes the conflicting productions. Additionally, U representing the utility of controlled production can be estimated as

$$U_i(n) = U_i(n-1) + \alpha[R_i(n) - U_i(n-1)]. \tag{2}$$

Here, α represents the learning rate, and $Ri(n)$ denotes the reward obtained by production i at time n. In general, rewards occur when a production associated with the goal of the task is executed. Typically, rewards are backpropagated to the productions that are executed before the reward is triggered. Each time a production is rewarded, the utility values of all productions that have been executed since the last update $(n-1)$ are updated using Equation 2. In this study, events related to intellectual curiosity and boredom are represented by assigning positive and negative rewards, respectively.

3.2.2 Reduction of pattern matching through production compilation

As mentioned earlier, production compilation compresses productions to reduce frequencies of pattern matching. Therefore, the fun generated by pattern matching (identifying structures in the data) was considered decayed by the compression accompanied with production compilation.

Figure 3 depicts the traces of an ACT-R model in a maze task used in simulations performed in this study. The vertical axis indicates time, and each column indicates an event in a module. The left-hand side trace represents the process of identifying the path from the declarative knowledge using pattern matching. The trace on the right-hand side expresses the search for a path without

[2] Claimed by Anderson (2007) in the chapter featuring "dynamic pattern matching" in ACT-R.

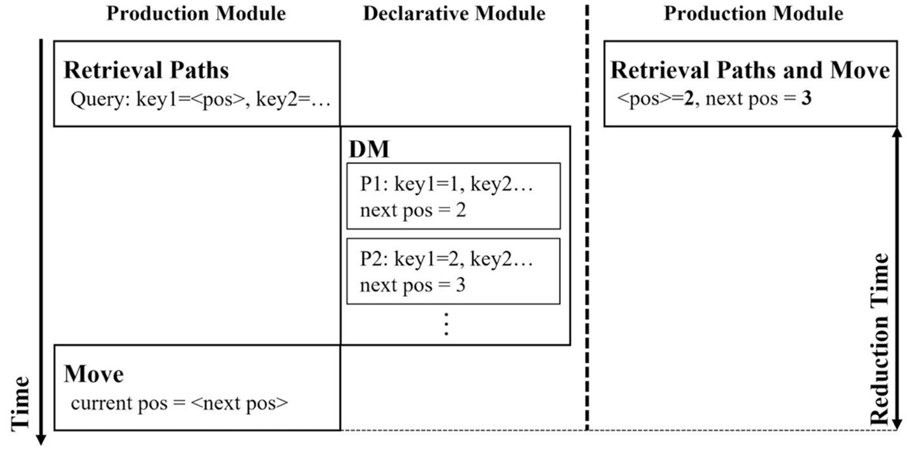


FIGURE 3
Example illustrating the before and after of learning using the production compilation.

pattern matching or retrieving paths from the declarative module; in other words, it represents the processing after production compilation.

3.3 Integrated mechanism of task continuation based on intellectual curiosity

We propose a mechanism for determining the continuation or termination of a task based on intellectual curiosity. Figure 4 illustrates the procedure of task continuation when executing general tasks. At the beginning of each round (unit related to the continuation of a task), the model determines whether to continue or terminate the task based on the conflict resolution between the two productions (*stop* and *continue productions*). The model proceeds with the round by firing various productions, such as searching the map, after deciding to continue the task. When the model encounters a condition that terminates the round, a new round is initiated, and the model again determines whether the task should be continued or terminated.

In the aforementioned process, the assigned initial values of utilities are higher in the continue production than in the stop production. At the beginning of the task, it can be assumed that agents intend to continue the task. The process of experiencing boredom from this initial state can be modeled by assigning a trigger of negative reward to the production recognizing the end of each round.[3] The utility of the production decreases when a negative reward is generated by the continue production at the end of the round, which in turn increases the firing probability of the stop production.

To deter boredom and continue the task, positive rewards corresponding to “fun” are necessary. This study associates the occurrence of pattern matching with the feeling of fun. We consider that this association is consistent with the definition of fun reported by Schmidhuber (2010) because it involves the discovery of patterns in the environment. However, repeated application of the same production causes habituation (production compression) and increases the opportunity to generate negative rewards to the continue production at the end of a round. In other words, the factor that ensures task continuation in the mechanism is the continued stimulation of intellectual curiosity through the discovery of declarative knowledge (data), which is the target of pattern matching.

4 Simulation

We performed simulations to verify the proposed mechanism of intellectual curiosity. This section explains the purpose of the simulations, the employed task, model details, and other settings involved in the simulations. Finally, the obtained results are summarized.[4]

4.1 Aims and indicators

To examine the mechanism of intellectual curiosity based on pattern discovery, we address the following questions.

1. What type of environment stimulates intellectual curiosity?
2. How does stimulated intellectual curiosity affect task learning?
3. What is the relationship between the proposed mechanism and the curiosity represented in existing reinforcement learning models?

The first question was answered by distinguishing between *external* and *internal* environments surrounding the model. Here, based on the previous discussion on IMRL (Singh et al., 2005), we adopted the term internal environment to explore the individual

[3] A similar mechanism of stopping a navigation task was presented by Anderson et al. (1993), although the exact equations of utility calculation were different because of the architectural difference.

[4] The models and data are included in https://github.com/AcmlNagashima/CuriosityAgents.

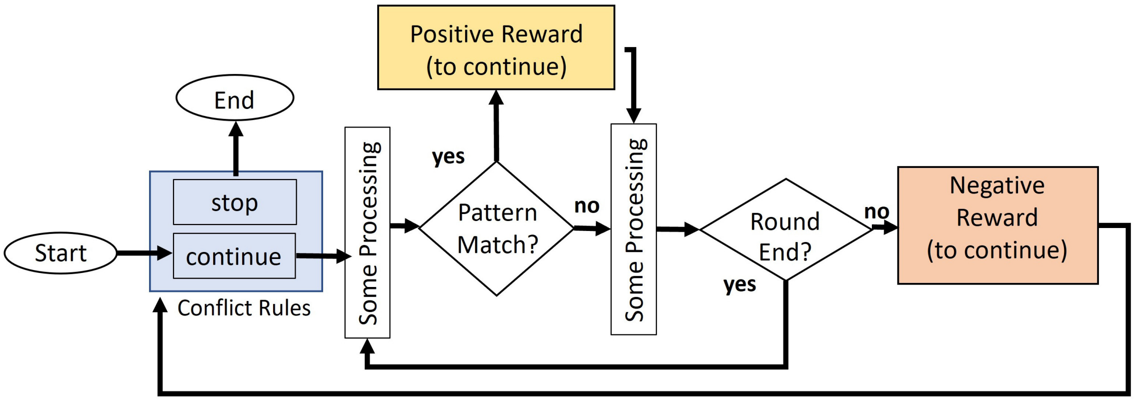


FIGURE 4
Flowchart of the task continuation model. The model generates positive rewards when pattern matching occurs.

differences surrounding and affecting intellectual curiosity. In this context, the external environment was manipulated by varying the complexity (difficulty level) of the learning environment, while the internal environment was defined as the strategy employed by the model to explore the external environment.

Furthermore, we examined how the factors of the internal and external environment affect intellectual curiosity using the indicators

(a) up-time ratio (percentage of time the model was running relative to the time limit of one run); and
(b) number of rounds (frequency of firing the task continuation production, depicted in Figure 4).

These indicators represent the extent to which the model engaged with the task. By definition, if the model obtains strong motivation, these indicators are assumed to be increased. We explored the internal and external environments that fostered this effect. As an internal environment, we manipulated the depth of thinking when searching external environments. According to the discussion presented in Section 2.1, this factor is expected to affect the effect of intrinsic motivation of the model via the interactions with the difficulty level of the external environment.

To answer the second question, the effect of stimulated intellectual curiosity on learning in the task was examined using the indicators

(c) entropy (variety of behavior patterns in the environment search);
(d) goal rate (the goal achievement rate); and
(e) the number of newly generated productions (frequency of occurrence of production compilation).

These indicators quantify the effect of intellectual curiosity on three aspects, namely, the behavior pattern (c), learning outcome (d), and internal states (e). We assumed that these indicators would increase with higher intellectual curiosity. In other words, the higher the motivation, the more opportunities the model has to explore the map. Moreover, as the model is extensively exploring the map, entropy (c) and the goal rates (d) increase while the model discovers more patterns in the external environment (e).

The complexity of model behavior (c) was computed as the entropy normalized for the frequency of occurrence of states of the task as follows:

$$Hr = \frac{-\sum_{i \in n} p(x_i) \log p(x_i)}{\log n} \tag{3}$$

Here, x_i denotes a particular state in an environment, and n represents the total number of states in an environment. This index increased when the model extensively explored the environment and the value decreased during local behaviors.

Finally, to address the last question, we used these indicators to examine whether the proposed behavior of curiosity was consistent with the previous models of curiosity. Among several models, we focused on the ICM model (Pathak et al., 2017) as the representative mechanism of deep reinforcement learning with curiosity and compared it with the ACT-R models with various internal and external environments.

4.2 Task: manipulation of the external environment

Based on the previous reports (Fu and Anderson, 2006; Reitter and Lebiere, 2010) on ACT-R explained in Section 2.3.2, we adopted the task of searching mazes. To systematically manipulate the difficulty level of the task, we applied a maze generation algorithm[5] to grids of sizes 5 × 5, 7 × 7, and 9 × 9, with 10 different maps prepared for each size; Figure 5 depicts an example of the maps. As indicated in the figure, the created mazes are loop-less structures with the starting location at the top-leftmost corner, and the goal location at the corner where the maximum number of corner points is traversed from the start point. In other words, two corner points with the highest number of hops were selected as the start and goal locations. The difficulty level of this task corresponded to the size of the maps. As described in Section 2.2, we assumed that an appropriate level of difficulty stimulates intrinsic motivation.

[5] https://algoful.com/Archive/Algorithm/MazeExtend.

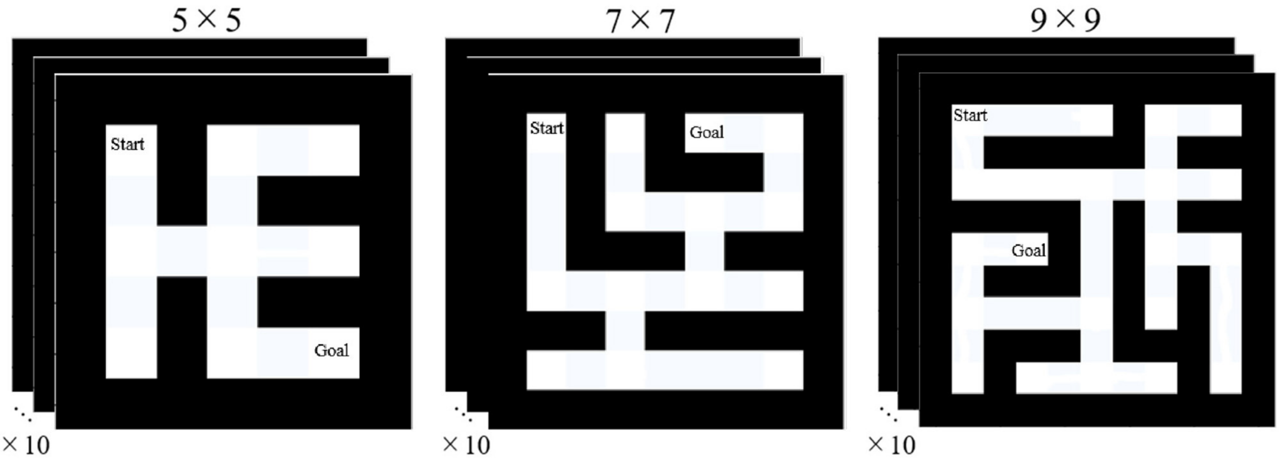


FIGURE 5
Manipulation of external environments.

Therefore, the factors that stimulate the proposed intellectual curiosity were examined by comparing different sizes of the external environments.

This task was implemented in ACT-R using a simplified method to obtain stable results over numerous runs. Rather than presenting a visual representation of the map to the model, we included chunks representing the structure of the map in the declarative module of the model.[6] In other words, the task corresponded to a situation where the model performed path planning without actually moving the body with respect to the topologically represented declarative knowledge of the environment.

The topological map provided to the model comprised chunks representing nodes (corner points) and paths (connections between the corner points) of the maze. When the task was executed, the model stored a node chunk in the goal module that indicated the currently focused corner point. From this state of the goal module, the model attempted to discover the chunk of paths stored in the declarative modules by matching them with patterns of variables embedded in the productions. When the chunk containing the current node was retrieved from the declarative module, the other node associated with the corresponding path chunk was newly stored in the goal module. This process was repeated until the model reached the goal point or the designated time was elapsed.

4.3 Search strategy: manipulation of the internal environment

To examine the internal environment that stimulates intellectual curiosity, we manipulated the strategy of exploring the external environment in terms of different levels of thinking (Brooks, 1986; Evans, 2003; Kahneman, 2011). As explained in Section 2.1, human mental activities are traditionally divided into at least two levels despite a continuous debate on the simple separation. This study follows the discussion reported by Conway-Smith and West (2022), suggesting that individual mental process is characterized by a spectrum between the fast automatic and slow deliberate processes. According to them, the levels in this spectrum can determine the amount of mental effort (computational cost) required for the task. Among several types of computational costs, we focused on the effort of retrieving declarative knowledge. As described in Section 2.3.1, retrieval of declarative knowledge in ACT-R can be hypothesized to increase prefrontal cortex activity. Therefore, it can be reasonably assumed that deliberative levels of thinking, which affect the optimal level of intrinsic motivation (Csikszentmihalyi, 1990; Yerkes and Dodson, 1908), are estimated from the amount of declarative knowledge retrieved during the task execution.

Figure 6 depicts the manipulation of the levels of thinking in this study focusing on the maze search task. The process of the model became complex from left to right, and the amount of declarative knowledge used in the task was assumed to increase. These models were developed based on the authors' previous work (Nagashima et al., 2021) with two modifications; more complex pattern matching in the path retrieval and leveraging all pattern matching as triggers of intrinsic reward. In the previous research, the smallest number of variables in the productions was only one, so there was no pattern in the rule. Also, the previous research limited the triggers of intrinsic rewards only when the maze searching rules were fired, omitting rewards generated from pattern matching that occurred by other productions during the task.

These changes were made to ensure the model's consistency with our theoretical assumptions. There may be debate about assuming that every pattern match triggers intrinsic rewards. For example, it might be possible to prioritize pattern matching based on complexity or to select productions for positive rewards by setting certain criteria. However, in this study, we prioritized a simpler setting to verify the basic idea, avoiding any arbitrariness. The next sections explain the specific process of the model in each internal environment.

[6] The exclusion of perceptual and motor processes in basic simulations is also recommended in the official ACT-R tutorial.

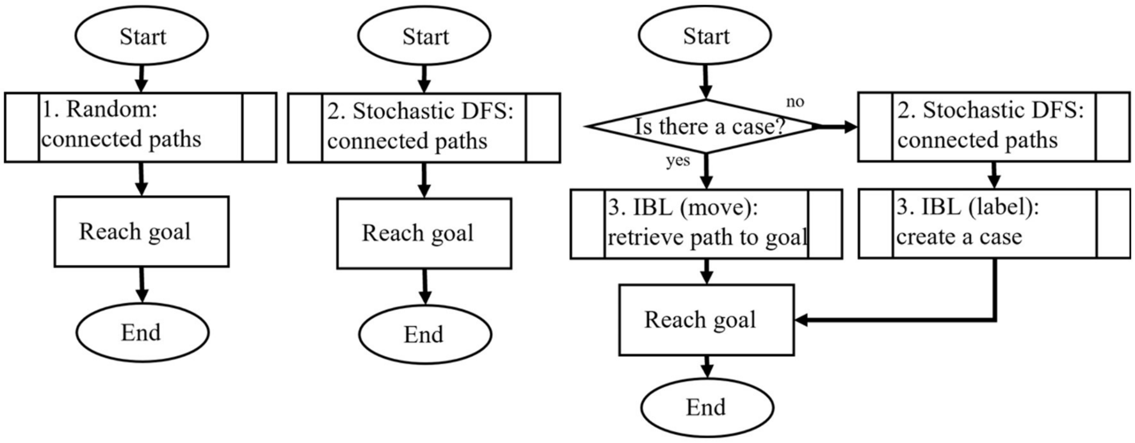

Start
1. Random: connected paths
Reach goal
End
Start
2. Stochastic DFS: connected paths
Reach goal
End
Start
Is there a case?
no
yes
2. Stochastic DFS: connected paths
3. IBL (move): retrieve path to goal
3. IBL (label): create a case
Reach goal
End

FIGURE 6
Manipulation of internal environments. DFS, depth-first search; IBL, instance-based learning.

4.3.1 Random model

The model with the lowest level of thinking randomly transitioned to the current location stored in the goal module. The model repeated the following process during each round until the goal was achieved or the time limit was reached.

1. Path search: the model retrieved the declarative knowledge related to the paths adjacent to the current location. To retrieve the declarative knowledge, the model used productions in which the current location was bound to a variable.
2. Move:
   (a) If the path retrieval was successful (pattern matching occurred), the model updated the state of the goal module according to the retrieved path, and the model returned to Step (1).
   (b) If the path retrieval failed, the model returned to Step (1) without modifying the state of the goal module.

While the model explored the maze using this search strategy, the productions that were used for the successful retrieval of the path were compiled. The model was assumed to have a few opportunities for pattern matching because production compilation occurred only when the stored path was retrieved. Therefore, stimulating intellectual curiosity in this model was considered as difficult.

4.3.2 Stochastic Depth-first Search (DFS) model

To include higher cognitive functions (declarative module), we constructed a probabilistic DFS model, which backtracked to search the environment based on the study by Reitter and Lebiere (2010). As indicated in Figure 7, the model exhibited a stacked structure with chunks generated by the imaginal module of ACT-R. The push function in the stack was realized by storing a chunk that contained the name of the previous chunk in the *Link* slot. Additionally, the pop function in the stack was realized by returning this slot value to the previous slot value. We implemented all these processes using only ACT-R productions without defining any external functions written in other programming languages, such as LISP.

Similar to the random model, the stochastic DFS model compiled productions that could retrieve declarative knowledge of paths and backtrack to learn new productions that did not contain variables. The specific model behavior can be summarized as follows.

1. Path search: the model determined the destination by retrieving the declarative knowledge associated with the path, similar to the random model. The IF clause included the current location stored in the goal buffer and five variables, corresponding to the current location and the directions (west, north, east, and south), which were flags indicating whether the direction was already searched or not.
2. Move:
   (a) If the knowledge retrieval was successful (pattern matching occurred), the model flagged the retrieved direction as "searched," created a new chunk using the imaginal module, and stored the chunk as declarative memory, as depicted in Figure 7. Simultaneously, the model updated the current location of the goal buffer according to the retrieved path. At this point, the searched flag in the goal buffer was reset, whereas the searched flag in the direction opposite to the direction of movement was set to prevent its return to the previous location. After this procedure, the model returned to Step (1).
   (b) The backtracking process was executed if the path retrieval failed, returning the model to the previous state by popping chunks in the stack; eventually, the model returned to Step (1).

The model repeated this behavior until the goal was achieved or the time limit was reached. Contrary to the random model, the DFS model used the stack when the path search failed. Therefore, the model required more rounds to compress (compile the production) the declarative knowledge of the paths.

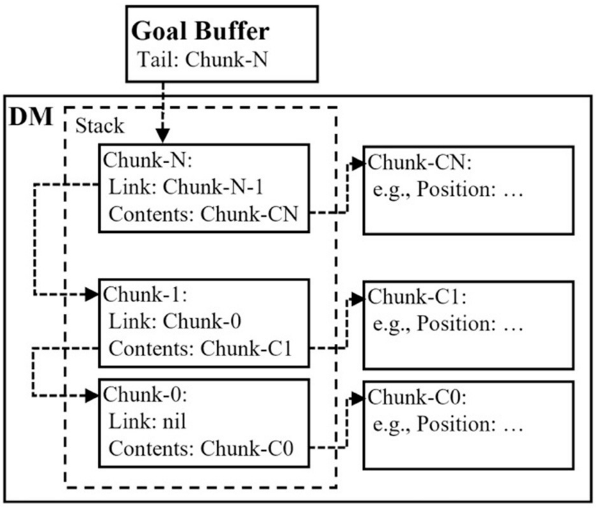


FIGURE 7
Construction of the stack structure using chunks of adaptive control of thought-rational (ACT-R) architecture. This stack was implemented using the imaginal module of ACT-R.

4.3.3 Stochastic DFS plus Instance-based Learning (IBL) model

This model combined the stochastic DFS with the IBL, which leverages past memories to solve current tasks (Gonzalez et al., 2003; Lebiere et al., 2007). In this task, the model held all the retrieved paths in the stack from the beginning of each round until the goal was reached. After the model attained the goal, the path chunks in the stack were retrieved one by one, and the chunks labeled "correct path" were generated. During each round, the model repeated the following two steps until the goal was achieved or the time limit was reached.

1. Determining strategies: at the beginning of each round, the model decided between the stochastic DFS and the IBL strategies by retrieving chunks associated with the current location and labeled "correct path."
2. Move:
   (a) When the DFS strategy was employed (failed to retrieve the "correct path"), the model behaved as a stochastic DFS model.
   (b) When the model successfully retrieved the "correct path," the model updated the current location according to the retrieved path chunk. Subsequently, the model returned to Step (1).

The model behaved as the stochastic DFS model in the early stages of the task. With the repetition of rounds and the increase in the number of instances with the "correct path," the model effectively reached the goal. Here, IBL was a time-consuming process in comparison with the DFS strategy. This was because the model had to retrieve the path in the stack at the end of the round to assign a label to a path. Moreover, retrieval trials for past successful rounds at the beginning of each round resulted in additional time, which was not included in the other models. We hypothesized that similar to the stochastic DFS model, this model is likely to stimulate intellectual curiosity, and the IBL function would positively affect the learning of the task.

4.3.4 Deep reinforcement learning model based on curiosity

To explore the relationship between the aforementioned ACT-R models and previous models of intrinsic motivation using deep reinforcement learning, we constructed an ICM model based on the report by Pathak et al. (2017). The ICM model in this study explored the maze using the policy π in actor-critic model.[7] This search resulted in a policy that maximized the rewards represented as

$$r_t = r_t^i + r_t^e. \tag{4}$$

Thus, the reward of the model was calculated as the sum of the internal reward (r_i) and external reward (r_e). Based on this equation, the model explored the environment by balancing the two types of rewards.

In this study, following Pathak et al. (2017), the internal reward was determined by

$$r_t^i = \frac{\eta}{2} \left\| \hat{\phi}\left(s_{t+1}\right) - \phi\left(s_{t+1}\right) \right\|_2^2 . \tag{5}$$

Here, state s is defined as pixel data in deep reinforcement learning. In this study, the maze situation (players, walls, and paths) was converted into a grayscale image (42 × 42) and served as input to a CNN, whose parameters were represented as ϕ. By subtracting the predicted and actual outputs of CNN, prediction errors were computed and weighted using the coefficient η. This coefficient was regarded as the intensity of curiosity.

By contrast, the external reward was defined as

$$r_e = \begin{cases} -1 & \text{if failed to move;} \\ 0 & \text{if succeeded to move;} \\ 10 & \text{if the goal was achieved.} \end{cases} \tag{6}$$

In each action, the model attempted to select one of the directions, namely, west, north, east, or south, and to transition the state from one corner point to another. If the model selected a direction that did not lead to a path, the action was considered a failure. If the model reached the goal owing to its movement, it was rewarded for its success; subsequently, the task moved on to the next round.

The search for the maze was terminated when the condition

$$th < r_i \times 500 + egs \tag{7}$$

was satisfied. Here, th denotes the threshold value, and egs indicates noise. The model search was terminated when the internal reward was less than the threshold.[8]

[7] Discount rate $gamma = 0.99$.

[8] A fixed value of 500 was tentatively multiplied because the scales of the internal reward in the ACT-R and ICM models were different.

4.4 Simulation settings

4.4.1 Setting for ACT-R models

As parameters relevant to the general model of task continuation (Figure 4), the simulation assigned the initial utility values of the *continue* and *stop productions* to 10 and 5, respectively. Additionally, we assigned the triggers of the negative reward ($r = 0$) to productions that recognized the end of the round, which was either reaching the goal or recognizing that the time limit of each round was elapsed. Conversely, the triggers of the positive reward were assigned to productions that included pattern matching, which corresponded to intellectual curiosity. We manipulated the intensity of the model's intellectual curiosity by sampling the positive reward at five equal intervals, ranging from 2 to 18. For parameters not directly related to our proposed mechanism, we adopted values from previous studies. Following Anderson et al. (2004), the activation noise level (ANS), which represents the noise in memory recall, was set to 0.4, and the production noise level (EGS), which reflects the noise in comparing utilities for continuing or terminating productions, was set to 0.5.

To enable the above setting of rewarding by pattern matching, we made small modifications to the original source code of ACT-R (Ver. 7.21). We first modified the source code of ACT-R to assign the reward trigger at any time point after the production compilation occurred. Subsequently, we modified the code to not inherit those triggers after the compilation. In the original ACT-R source code, the compiled production inherits the reward triggers from the original production. We redefined this function to represent boredom caused by the lack of new production compilation.

Simulations based on these settings were run 10 times for each map and each positive reward setting. The limits in the ACT-R simulation time for each round and run were set to 180 and 3,600 s, respectively.

4.4.2 Setting for ICM model

The ICM model was implemented using PyTorch (ver. 1.9.0), with parameters set to match those of the ACT-R, wherein the simulations were run on 30 maps and the proportion of the internal reward (η) for each run was divided into five samples with equal intervals, ranging from 0.1 to 0.9. We compared the sum of the internal reward (r_i) and the noise (*egs*) with the threshold ($th = 5$) in Equation 7 to determine whether the task was continued or terminated. Furthermore, we set 156 and 3,130 steps as the limit of the action in each round and run, respectively. These steps were set to be equivalent to the time limit set at the ACT-R models. One step of the ICM model was equivalent to the rule transition time of 1.15 s in the default random model. The ICM model was run 100 times for each reward setting as it could run faster than the ACT-R models.

4.5 Simulation results

Figures 8, 9 illustrate the simulation results as a function of the internal reward for each of the indicators discussed in Section 4.1. Each graph depicts the average value, which was $n = 100$ (10 times $\times$ 10 maps) for the ACT-R models and $n = 1,000$ (100 times $\times$ 10 maps) for the ICM model, aggregated for each internal and external environment condition with respect to the map size. The influence of the maps of the external environment was examined by comparing the three series in each graph, whereas the influence of the internal environment (random, DFS, DFS + IBL) of the model was analyzed based on the difference between the graphs aligned in the horizontal direction. The subsequent sections discuss the obtained results based on the three questions posed as objectives of the simulation.

4.5.1 Environment that stimulates intellectual curiosity

In Section 4.1, the first question posed was "What type of environment stimulates intellectual curiosity?" To address this question, we focused on the up-time ratio (Figure 8A) and number of rounds (Figure 8B) as the behavior indicators of intellectual curiosity. These indicators increased continuously with the increase in the strength of intellectual curiosity for every series (map size) in every graph (levels of thinking) of Figure 8. This general trend suggested that the implemented intellectual curiosity actually enhanced the motivation for the task and ensured task continuation.

In terms of the difference in the external environment, the up-time ratio (Figure 8A) increased as the map became more complex ($9 \times 9 > 7 \times 7 > 5 \times 5$). However, the number of rounds (Figure 8B) presented a reverse trend, wherein the simpler external environment increased the number of rounds ($9 \times 9 < 7 \times 7 < 5 \times 5$), except for the DFS model (Figure 8B2). The discrepancy between the two indices of motivation was caused by the time limit of the simulation (3,600 s). The model could complete the simple map faster, resulting in a greater number of rounds within the time limit (Figure 8B). However, as indicated by Figure 8A, the simple map enabled the model to terminate the task early owing to the lack of new patterns for production compilation. This result implies that the complex external environment stimulates intellectual curiosity.

Furthermore, we determined the difference between the internal environment, which was not expected in advance. When comparing the three horizontally aligned ACT-R models, we observed that the models with high levels of thinking (DFS + IBL and DFS) had smaller indicators of motivation than the random model. The reason for this difference could be the advantage of the random model with less thinking time and more trials and errors. In this condition without physical constraints, the random model had a better chance of receiving positive rewards by identifying novel paths than the other models.

4.5.2 Effect of task continuation on model learning

The second question posed was "How does stimulated intellectual curiosity affect task learning?" Figure 9 presents the three learning indices, namely, the changes in behavior (Figure 9C: entropy), the learning outcome (Figure 9D: goal rate), and the changes in internal state (Figure 9E: the number of productions).

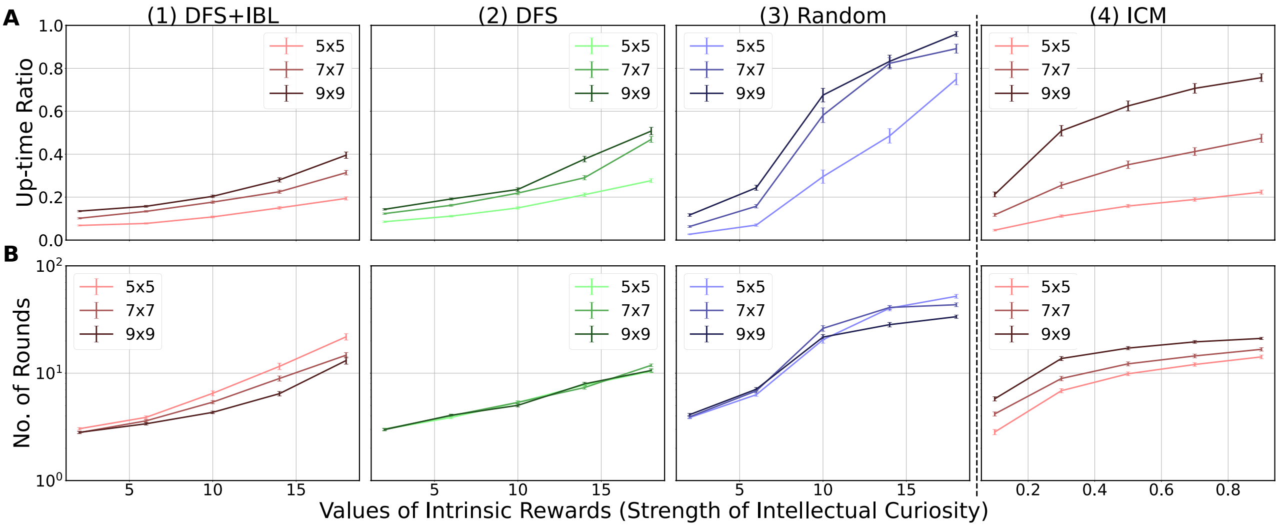


FIGURE 8
Simulation results. The numbers in the horizontal line distinguish the models (1–3: adaptive control of thought-rational (ACT-R) models; 4: intrinsic curiosity module (ICM) model), and the vertical alphabet differentiates the indicators (**A**: Up-time ratio; **B**: Number of rounds). The error bars in each graph indicate the mean value ($n = 10$) of the standard deviations (ACT-R: $n = 10$, ICM: $n = 100$) obtained for each map when multiplied by 1/10.

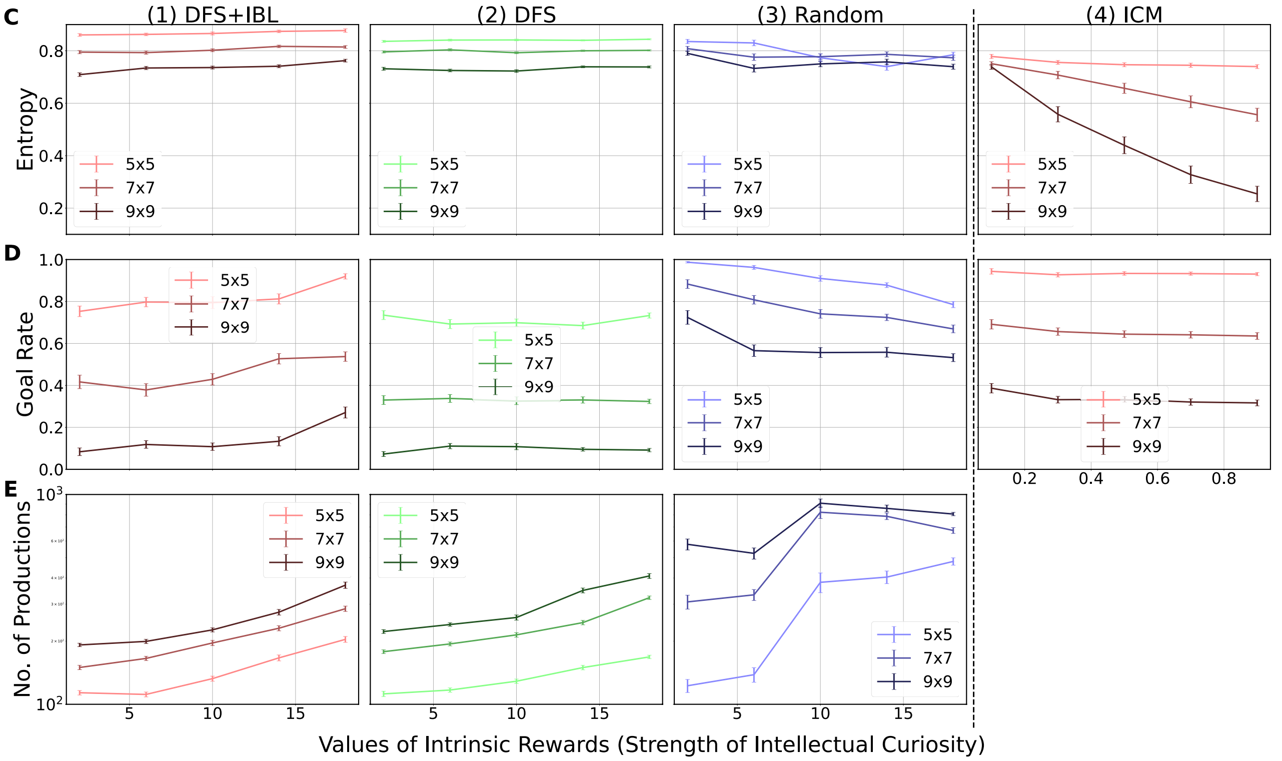


FIGURE 9
Simulation results. The numbers in the horizontal line distinguish the models (1–3: adaptive control of thought-rational (ACT-R) models; 4: intrinsic curiosity module (ICM) model), and the vertical alphabet differentiates the indicators (**C**: Entropy; **D**: Goal rate; **E**: Number of productions). The error bars in each graph indicate the mean value ($n = 10$) of the standard deviations (ACT-R: $n = 10$, ICM: $n = 100$) obtained for each map when multiplied by 1/10.

Based on the analysis of Figure 8, we confirmed that all conditions of the internal and external environments were stimulated by intellectual curiosity. However, Figure 9 indicates that the effect of intellectual curiosity on task learning differs depending on the internal environment. The intensity of intellectual curiosity affected positively for higher levels of thinking.

In the highest level of thinking (the DFS + IBL model), all learning indices (Figures 9C1, D1, E1) increased with the intensity of intellectual curiosity. In the middle level (the DFS model) increased the number of productions (Figure 9C2) while maintaining the entropy (Figure 9C2) and goal rate (Figure 9D2). In the case of the lowest level (the random model), the intrinsic motivation decreased all indices (Figures 9C3, D3, E3). These trends indicated that the DFS + IBL model exhibited a goal-oriented behavior because of the learning effect of the IBL strategy, whereas the behavior of the DFS model had to search the entire map. Furthermore, the random model did not lead to the goal; this was because the model reinforced unfavorable behavior by repeatedly visiting the same location without expanding the search.

In terms of the effect of the challenge of the task (task difficulty), the entropy (Figure 9C) and the goal rate (Figure 9D) were greater on the small map, whereas the number of productions was higher on the large map. These differences may be attributed to the fact that the small map was easier to explore, which in turn increased the entropy and the goal rate. Conversely, the large map exhibited more pattern-matching opportunities, leading to more accumulated knowledge by frequent compilation.

In summary, intellectual curiosity promoted learning in the DFS + IBL model, which exhibited the highest level of thinking. By contrast, learning in the DFS and random models was not promoted by intellectual curiosity. Moreover, the effect of intellectual curiosity negatively impacted the learning environment in the random model.

4.5.3 ACT-R curiosity vs. ICM curiosity

Finally, we compared the ICM and ACT-R models in Figures 8, 9. Similar to all ACT-R models, the ICM model was stimulated by stronger intellectual curiosity (Figure 8). However, the effect of intellectual curiosity for task learning was specifically similar to the random ACT-R model that exhibited decreasing trends of the entropy (Figure 9C4) and the goal rate (Figure 9D4) with the increase in the strength of intellectual curiosity. With respect to the effect of the external environment, the ICM model was also similar to the random ACT-R model; the up-time ratio (Figure 8A4) and the number of rounds (Figure 8B4) were greater for larger maps, whereas the entropy (Figure 9C4) and the goal rate (Figure 9D4) were greater for smaller maps.

This comparison confirmed commonalities and differences between the developed ACT-R curiosity model and the existing curiosity model in deep reinforcement learning. The proposed ACT-R curiosity mechanism can represent similar learning to the existing model by including a simple internal environment (random search strategy). At the same time, it can incorporate goal-directed behavior by including "explicit use of success memory." Such an explicit nature of the proposed mechanism also leads to a direct examination of the model's internal learning. The analysis of Figure 9E clearly shows this advantage of interpretability made by the proposed approach.

4.5.4 Cases of paths discovered by the models

To compare detailed behaviors between models, Figure 10 illustrates example paths in a 5 × 5 map. The map depicts start and goal positions at the top left and bottom right corners respectively. The circles' colors and line thickness represent visit frequencies during runs. The random model exhibited diagonal movement and movement through walls, a result of compiling multiple movement rules. To gather these examples, we conducted 10 runs for each model across three levels of intrinsic rewards, selecting the runs with the lowest and highest performance for analysis.

These figures reveal distinct behavioral characteristics of each model. The random model predominantly exhibits localized movements within specific areas, often distant from the goal. On the other hand, the DFS model explores the map evenly but does not necessarily move directly toward the goal. In contrast, the DFS + IBL model demonstrates deliberate behaviors aimed at reaching the goal, particularly under high-reward conditions. In terms of localized movement patterns, the ICM model was more similar to the random and DFS models than the DFS + IBL model. Thus, the results suggested that the DFS + IBL model had a greater effect on curiosity strength than the other models regarding directionality toward the goal. These observations support the findings observed in the quantitative results shown in Figures 8, 9.

5 Conclusions

The objective of this study was to develop a mechanism for intrinsic motivation based on pattern discovery by combining basic modules of ACT-R. This section summarizes the significance of the proposed mechanism and presents the potential future lines of investigation.

5.1 Summary and implications

The proposed mechanism was based on the assumption that pattern discovery is associated with the feeling of fun and is the source of intellectual curiosity. Additionally, its attenuation was expressed by the learning mechanism incorporated in ACT-R. To support this proposal, we implemented multiple external environments (challenges in the task) and strategies for exploring the external environment (levels of thinking) and examined the role of intellectual curiosity in each situation. The simulation results indicated that the rewards associated with pattern discovery exhibited different effects on models at different levels of thinking. The model with the lowest level of thinking (random) and that with the middle level of thinking (DFS) had negative and neutral effects of intellectual curiosity on performance, respectively. The only model that benefited from intellectual curiosity was the one with the highest level of thinking (DFS + IBL), which comprised a function that enabled it to remember previous experiences that led to the goal.

These results are partially consistent with the past arguments made for human intrinsic motivation. Particularly, the effectiveness of intrinsic motivation in the DFS + IBL model is consistent with a discussion, in which intrinsic motivation operates well with deliberative thinking, which requires "autonomy," "mastery," and "purpose" (Pink, 2011). Furthermore, consistent with our negative results in the random model, several reports exist on behavioral addictions caused by the negative effects of intrinsic motivation

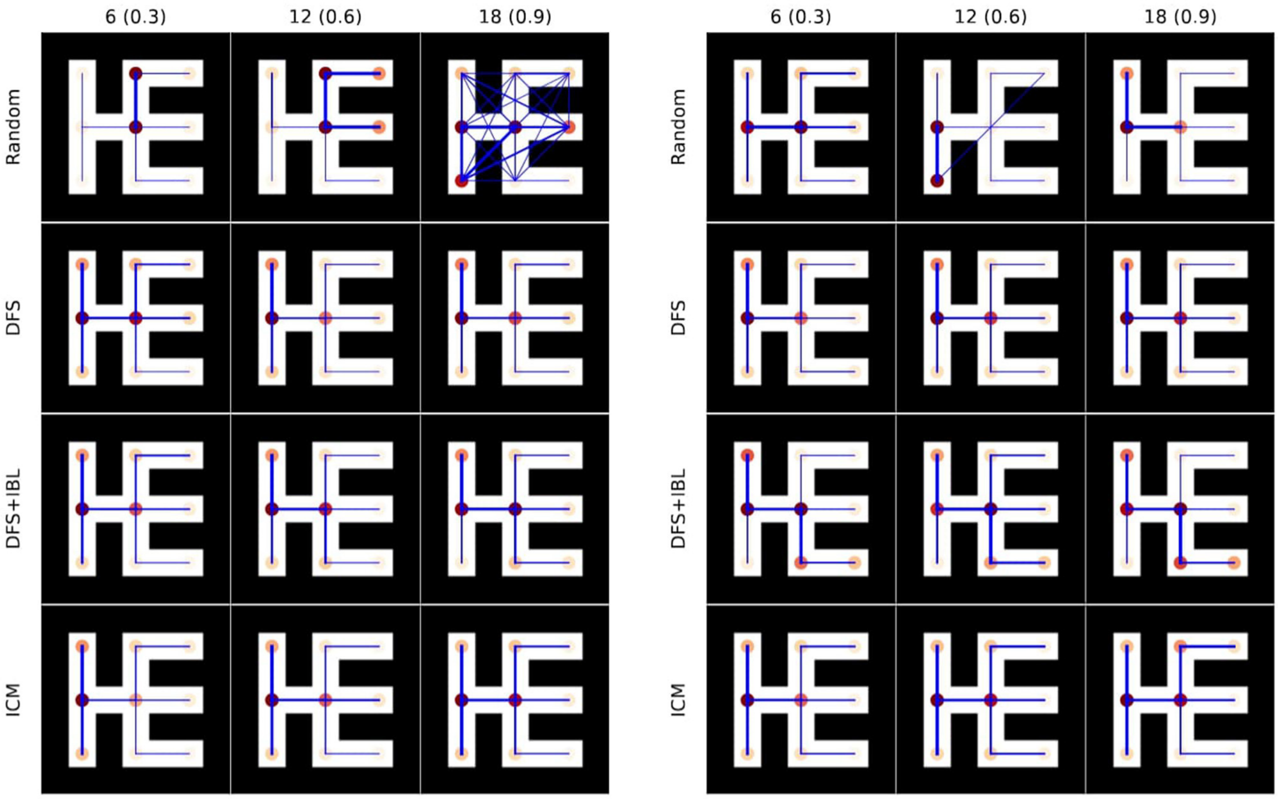


FIGURE 10
Trajectories of the model runs [**(left)**: low-performance runs, **(right)**: high-performance runs]. The columns indicate the strength of intellectual curiosity, while the rows correspond to each model.

(Alter, 2017). For instance, people often forget their goals and become engrossed in exploratory tasks, such as browsing the internet, resulting in poor performance. This irrational behavior might also relate to *computational psychiatry* (Huys et al., 2016).

In addition to the above discussion on the internal environment, we identified a connection with a previous discussion on the external environment. Consistent with the discussion on "challenge" (Malone, 1981), we determined that the up-time ratio in larger maps was greater than that in the smaller maps. This result indicates that a complex external environment stimulates intellectual curiosity. However, we also determined that difficult challenges generate ineffective learning on a wide map (Figure 9). The aforementioned positive and negative effects of the task difficulty indicate the optimal level of challenge (Csikszentmihalyi, 1990; Yerkes and Dodson, 1908).

Furthermore, this study successfully corresponded with past studies on intrinsic motivation in reinforcement learning. The comparisons with the ICM model (Pathak et al., 2017) confirmed that the developed ACT-R model, specifically the random model, is a succession of existing studies. Although we cannot claim its superiority as a learning algorithm based on the current simulation alone, the model with the higher level of thinking (DFS + IBL) exhibited characteristic behavior toward the ICM model. Future analysis of more extensively manipulating parameters, such as the balancing of r_i and r_e in Equation 4 and designing the external environment stimulating curiosity (Burda et al., 2018), could reveal further correspondence between the ACT-R model and reinforcement learning framework.

We believe that the comparisons of the previous model of reinforcement learning reveal the methodological advantage of using cognitive architecture. An integrated cognitive architecture, such as ACT-R, provides criteria to set numerical parameters (e.g., time limits and utilities) based on previous studies. Furthermore, ACT-R comprises neuroscientific modules that correspond to basic cognitive functions, such as declarative and procedural knowledge. Based on this relation, arguments associated with human intrinsic motivation can be developed. Therefore, this study contributes to the understanding of intrinsic motivation in a wide context of the relationship between human evolution and the development of civilization by mapping the discovery of patterns to intrinsic motivation (Baron-Cohen, 2020).

5.2 Future work

The proposed mechanism of intellectual curiosity has the potential for several future studies. The primary focus among them is human experiments that manipulate the internal and external environments as in the simulation. A simulation study without data is nothing more than a demonstration derived deductively from theory. Therefore, the model's value must be proven by applying it to human scenarios.

One of the obstacles to conducting human experiments for the proposed mechanism is setting tasks to stimulate human curiosity. In this study, we adopted the maze task because several previous researchers based on ACT-R have constructed models for this task. However, setting experimental situations with human participants to exhibit intrinsic motivation for solving such simple tasks may be difficult. Therefore, in the future, we intend to explore tasks that both humans and developed models can execute with proper intrinsic motivation.

Other future work will focus on modeling the curiosity and motivation that was not explained in the current study. As we discussed in Section 3.1, this study targeted on intellectual curiosity relating "a desire to bring better form to one's knowledge structures" (Malone, 1981) or "intrinsic desire to build a better model for the world" (Schmidhuber, 2010). Therefore, we have not yet explained the sensory curiosity that drives us to acquire new knowledge from the world. These two types of curiosity are considered complementary, similar to the explore-exploit trade-off in reinforcement learning. Without including sensory curiosity in the model, we cannot explain how declarative knowledge is acquired for intellectual curiosity, nor how the initial utility settings of continuing the task exceed those of stopping.

The above future study possibly leads to a deeper exploration of levels of thinking. Conway-Smith et al. (2023) recently summarized the relationship between metacognition and levels of thinking, arguing that compilation of existing knowledge reduces the effort involved in metacognition, making it more automatic. Building on this, we can suggest that achieving such a metacognitive state as a result of higher-level thinking enabled with enough intellectual curiosity, exemplified by the DFS + IBL model.

On the contrary, we can assume the exploratory role of the lower level of thinking. As shown in Figure 9, the random model showed a higher goal ratio with larger learning products in some conditions. These results suggest links between low-level thinking and sensory curiosity, leading to exploration of the environment. Our recent work (Nagashima and Morita, 2024) provides support for the above interpretation. In the experiment, human participants observed the behaviors generated by the models in the current study and rated the random model as having the most curious features.

The final direction for future research is the generalization of the ideas presented in this paper to other tasks in real-world settings. We believe such tasks are linked to the earlier discussion on the civilization of society (Baron-Cohen, 2020). As suggested by Toya and Hashimoto (2018), tool-making requires recursive compilation of intermediate products. Integrating this multi-agent simulation with the mechanisms proposed in the current study could offer a detailed explanation of the driving forces behind the evolution of civilization.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

KN: Software, Writing – original draft, Writing – review & editing. JM: Conceptualization, Writing – original draft, Writing – review & editing. YT: Conceptualization, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by JSPS KAKENHI Grant Number JP24KJ1215.

Acknowledgments

We acknowledge the use of ChatGPT-4.0 (Open AI, https://chat.openai.com/), DeepL (https://www.deepl.com/), and Grammarly (https://app.grammarly.com/) for assistance in editing and proofreading for this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2024.1397860/full#supplementary-material

References

Alter, A. (2017). *Irresistible: The Rise of Addictive Technology and The Business of Keeping Us Hooked.* London: Penguin.

Anderson, J. R. (2007). *How Can the Human Mind Occur in The Physical Universe.* New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780195324259.001.0001

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., Qin, Y., et al. (2004). An integrated theory of the mind. *Psychol. Rev.* 111, 1036–1060. doi: 10.1037/0033-295X.111.4.1036

Anderson, J. R., Kushmerick, N., and Lebiere, C. (1993). "Navigation and conflict resolution," in *Rules of The Mind* (Psychology Press), 93–120.

Atashfeshan, N., and Razavi, H. (2017). Determination of the proper rest time for a cyclic mental task using ACT-R architecture. *Hum. Factors* 59, 299–313. doi: 10.1177/0018720816670767

Aubret, A., Matignon, L., and Hassas, S. (2019). A survey on intrinsic motivation in reinforcement learning. *arXiv* [Preprint]. arXiv:1908.06976. doi: 10.48550/arXiv.1908.06976

Balaji, B., Shahab, M. A., Srinivasan, B., and Srinivasan, R. (2023). ACT-R based human digital twin to enhance operators' performance in process industries. *Front. Hum. Neurosci.* 17:18. doi: 10.3389/fnhum.2023.1038060

Baron-Cohen, S. (2020). *The Pattern Seekers: How Autism Drives Human Invention*. New York, NY: Basic Books.

Barrett, L. F. (2017). *How Emotions Are Made: The Secret Life of the Brain*. London: Pan Macmillan.

Bellemare, M. G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. (2016). "Unifying count-based exploration and intrinsic motivation," in *Proceedings of the 30th International Conference on Neural Information Processing Systems* (Barcelona), 1479–1487.

Bothell, D. (2020). *ACT-R 7.21+ Reference Manual*. Available at: http://act-r.psy.cmu.edu/actr7.21/reference-manual.pdf

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE J. Robot. Autom.* 2, 14–23. doi: 10.1109/JRA.1986.1087032

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., Efros, A. A., et al. (2018). Large-scale study of curiosity-driven learning. *arXiv* [Preprint]. arXiv:1808.04355. doi: 10.48550/arXiv.1808.04355

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. (2019). "Exploration by random network distillation," in *7th International Conference on Learning Representations (ICLR 2019)* (New Orleans, LA), 1–17.

Caillois, R. (1958). *Les Jeux et les Hommes: Le Masque et la Vertige*. Paris: Gallimard.

Ceballos, J. M., Stocco, A., and Prat, C. S. (2020). The role of basal ganglia reinforcement learning in lexical ambiguity resolution. *Top. Cogn. Sci.* 12, 402–416. doi: 10.1111/tops.12488

Conway-Smith, B., West, R. L., and Mylopoulos, M. (2023). "Metacognitive skill: how it is acquired," in *Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 45* (Sydney, NSW).

Conway-Smith, K., and West, R. L. (2022). "Clarifying system 1 & 2 through the common model of cognition," in *Proceedings of the 20th International Conference on Cognitive Modelling* (Toronto, ON), 40–45.

Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*. New York, NY: Harper & Row.

Damasio, A. R. (2003). *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*. Boston, MA: Houghton Mifflin Harcourt.

Dancy, C. L., Ritter, F. E., Berry, K. A., and Klein, L. C. (2015). Using a cognitive architecture with a physiological substrate to represent effects of a psychological stressor on cognition. *Comput. Math. Organ. Theory* 21, 90–114. doi: 10.1007/s10588-014-9178-1

Evans, J. S. (2003). In two minds: dual-process accounts of reasoning. *Trends Cogn. Sci.* 7, 454–459. doi: 10.1016/j.tics.2003.08.012

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Fu, W., and Anderson, J. R. (2006). From recurrent choice to skill learning: a reinforcement-learning model. *J. Exp. Psychol. Gen.* 135, 184–206. doi: 10.1037/0096-3445.135.2.184

Gonzalez, C., Lerch, J. F., and Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cogn. Sci.* 27, 591–635. doi: 10.1207/s15516709cog2704_2

Gunzelmann, G., Byrne, M. D., Gluck, K. A., and Moore Jr, L. R. (2009). Using computational cognitive modeling to predict dual-task performance with sleep deprivation. *Hum. Factors* 51, 251–260. doi: 10.1177/0018720809334592

Huizinga, J. (1939). *Homo Ludens Versuch einer Bestimmung des Spielelementest der Kultur*. Amsterdam: Pantheon.

Huys, Q. J., Maia, T. V., and Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* 19, 404–413. doi: 10.1038/nn.4238

Juvina, I., Larue, O., and Hough, A. (2018). Modeling valuation and core affect in a cognitive architecture: the impact of valence and arousal on memory and decision-making. *Cogn. Syst. Res.* 48, 4–24. doi: 10.1016/j.cogsys.2017.06.002

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Macmillan.

Koster, R. (2013). *Theory of Fun for Game Design*. Sebastopol, CA: O'Reilly Media.

Kotseruba, I., and Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artif. Intell. Rev.* 53, 17–94. doi: 10.1007/s10462-018-9646-y

Laird, J. E., Lebiere, C., and Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Mag.* 38, 13–26. doi: 10.1609/aimag.v38i4.2744

Lebiere, C., Gonzalez, C., and Martin, M. (2007). "Instance-based decision making model of repeated binary choice," in *Proceedings of the 8th International Conference on Cognitive Modelling* (Ann Arbor, MI), 67–72.

LeDoux, J. E., Pine, D. S. (2016). Using neuroscience to help understand fear and anxiety: a two-system framework. *Am. J. Psychiatry* 173, 1083–1093. doi: 10.1176/appi.ajp.2016.16030353

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cogn. Sci.* 5, 333–369. doi: 10.1016/S0364-0213(81)80017-1

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., et al. (2016). "Asynchronous methods for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning* (New York, NY: PMLR), 1928–1937.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. doi: 10.1038/nature14236

Nagashima, K., and Morita, J. (2024). "Trait inference on cognitive model of curiosity: relationship between perceived intelligence and levels of processing," in *Proceedings of the 22th International Conference on Cognitive Modelling* (Tilburg).

Nagashima, K., Morita, J., and Takeuchi, Y. (2021). "Curiosity as pattern matching: Simulating the effects of intrinsic rewards on the levels of processing," in *Proceedings of the 19th International Conference on Cognitive Modelling*, 197–203.

Nagashima, K., Nishikawa, J., Yoneda, R., Morita, J., and Terada, T. (2022). "Modeling optimal arousal by integrating basic cognitive components," in *Proceedings of the 20th International Conference on Cognitive Modeling* (Toronto, ON), 196–202.

Nishikawa, J., Nagashima, K., Yoneda, R., Morita, J., and Terada, T. (2022). "Representing motivation in a simple perceptual and motor coordination task based on a goal activation mechanism," in *Advances in Cognitive Systems 2022 (ACS 2022)* (Chicago, IL), 102–120.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). "Curiosity-driven exploration by self-supervised prediction," in *In Proceedings of the 34th International Conference on Machine Learning* (Sydney, NSW: PMLR), 2778–2787. doi: 10.1109/CVPRW.2017.70

Pink, D. H. (2011). *Drive: The Surprising Truth about What Motivates Us*. London: Penguin.

Raffaelli, Q., Mills, C., and Christoff, K. (2018). The knowns and unknowns of boredom: a review of the literature. *Exp. Brain Res.* 236, 2451–2462. doi: 10.1007/s00221-017-4922-7

Reitter, D., and Lebiere, C. (2010). A cognitive model of spatial path-planning. *Comput. Math. Organ. Theory* 16, 220–245. doi: 10.1007/s10588-010-9073-3

Ritter, F. E., Tehranchi, F., and Oury, J. D. (2019). ACT-R: a cognitive architecture for modeling cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* 10:e1488. doi: 10.1002/wcs.1488

Rosenbloom, P., Laird, J., Lebiere, C., Stocco, A., Granger, R., Huyck, C., et al. (2024). "A proposal for extending the common model of cognition to emotion," in *Proceedings of the 22th International Conference on Cognitive Modeling* (Tilburg).

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Trans. Auton. Ment. Dev.* 2, 230–247. doi: 10.1109/TAMD.2010.2056368

Singh, S., Barto, A. G., and Chentanez, N. (2005). "Intrinsically motivated reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning* (Sydney, NSW: PMLR), 2778–2787. doi: 10.21236/ADA440280

Spiro, R. J., Feltovich, P. J., Jacobson, M. J., and Coulson, R. L. (2012). "Cognitive flexibility, constructivism, and hypertext: random access instruction for advanced knowledge acquisition in ill-structured domains," in *Constructivism in Education*, eds. L. P. Steffe, and J. Gale (New York, NY: Routledge), 85–107.

Stocco, A., Sibert, C., Steine-Hanson, Z., Koh, N., Laird, J. E., Lebiere, C. J., et al. (2021). Analysis of the human connectome data supports the notion of a "Common Model of Cognition" for human and human-like intelligence across domains. *Neuroimage* 235:118035. doi: 10.1016/j.neuroimage.2021.118035

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge: MIT Press. doi: 10.1109/TNN.1998.712192

Taatgen, N. A., and Lee, F. J. (2003). Production compilation: a simple mechanism to model complex skill acquisition. *Hum. Factors* 45, 61–76. doi: 10.1518/hfes.45.1.61.27224

Toya, G., and Hashimoto, T. (2018). Recursive combination has adaptability in diversifiability of production and material culture. *Front. Psychol.* 9:1512. doi: 10.3389/fpsyg.2018.01512

van der Velde, M., Sense, F., Borst, J. P., van Maanen, L., and Van Rijn, H. (2022). Capturing dynamic performance in a cognitive model: estimating ACT-R memory parameters with the linear ballistic accumulator. *Top. Cogn. Sci.* 14, 889–903. doi: 10.1111/tops.12614 van Vugt, M. K., and van der Velde, M. (2018). How does rumination impact cognition? a first mechanistic model. *Top. Cogn. Sci.* 10, 175–191. doi: 10.1111/tops.12318

Xu, Y., and Stocco, A. (2021). Recovering reliable idiographic biological parameters from noisy behavioral data: the case of basal ganglia indices in the probabilistic selection task. *Comput. Brain Behav.* 4, 318–334. doi: 10.1007/s42113-021-00102-5

Yang, Y. C., and Stocco, A. (2024). Allocating mental effort in cognitive tasks: a model of motivation in the ACT-R cognitive architecture. *Top. Cogn. Sci.* 16, 74–91. doi: 10.1111/tops.12711

Yerkes, R. M., and Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *J. Comp. Neurol. Psychol.* 18, 459–482. doi: 10.1002/cne.920180503

frontiers | Frontiers in Psychology

TYPE Opinion
PUBLISHED 05 November 2024
DOI 10.3389/fpsyg.2024.1410462

OPEN ACCESS

EDITED BY
Simone Belli,
Complutense University of Madrid, Spain

REVIEWED BY
Meisam Dastani,
Gonabad University of Medical Sciences, Iran
Grant Bollmer,
University of Maryland, United States

*CORRESPONDENCE
Jie Wu
✉ wujie98@ruc.edu.cn

RECEIVED 01 April 2024
ACCEPTED 21 October 2024
PUBLISHED 05 November 2024

CITATION
Wu J (2024) Social and ethical impact of emotional AI advancement: the rise of pseudo-intimacy relationships and challenges in human interactions.
Front. Psychol. 15:1410462.
doi: 10.3389/fpsyg.2024.1410462

Social and ethical impact of emotional AI advancement: the rise of pseudo-intimacy relationships and challenges in human interactions

Jie Wu*

School of Journalism and Communication, Renmin University of China, Beijing, China

KEYWORDS

emotion, algorithm, user-platform relationship, pseudo-intimacy relationship, paradigm for human interaction

1 Introduction

Algorithmic platforms, as a form of intelligent digital infrastructure underpinned by algorithms, have fundamentally transformed how individuals connect and interact with one another (Yin and Lin, 2023). The integration of emotional intelligence into these algorithms further deepens the relational connections between users and platforms. By enhancing the algorithm's affective capabilities—such as emotion perception and feedback (Wu et al., 2022; Bie and Zeng, 2024; Peng, 2024)—the attributes of user-platform relationships undergo qualitative changes within the affective dimension. Consequently, the model of human-computer interaction (HCI) evolves toward a trend of humanization, as articulated by Paul (2017). To some extent, this not only realizes the technical possibility of platform personification but also addresses modern individuals' emotional needs within Cyborg space. This evolution fosters and sustains potential connections in the emotional dimension between users and platforms (Lai, 2023; Hong and Huang, 2024). As a result, emotions have emerged as a significant focus in research concerning interactions between users and algorithmic platforms.

Going back in history, Marvin Minsky, regarded as the father of artificial intelligence (AI), proposed the groundbreaking idea that "AI should possess emotions" as early as 1985 (Marvin, 2006). Subsequently, Rosalind Picard, in her seminal work *Affective Computing*, further elucidated the technical possibilities of endowing computers with emotional capabilities (Rosalind, 1997). Numerous studies in media psychology have demonstrated that individuals often mindlessly equate virtual media featuring anthropomorphic cues with real-life experiences (Reeves and Nass, 1996), leading to para-social interactions and relationships with these entities (Rubin et al., 1985; Bickmore and Picard, 2005). However, at that time, this conclusion was constrained by limitations within computer science and artificial intelligence fields regarding affective computing technology. Consequently, it primarily existed at the level of academic experiments and discussions without widespread empirical evidence in real-world contexts. In recent years, however, the rapid advancement and extensive application of emotional AI have transcended these academic boundaries. This evolution has led to para-social relationships becoming increasingly normalized within society—an occurrence that is now garnering significant attention from a diverse array of scholars across the humanities and social sciences.

Unlike scholars in the field of affective computing, who primarily focus on experimental research involving HCI, those in the humanities and social sciences tend to emphasize the exploration of the social and ethical risks associated with these technologies from philosophical and sociological perspectives. For instance, Marx famously asserted that the essence of humanity is "the sum of all social relations" (Marx Engels, 2009). Consequently, understanding what it means to be human necessitates an examination of the relationships between individuals and others. Building upon this foundation, some scholars have suggested that within the current dynamics of user-platform interactions, integrating emotional intelligence into the development of human-like AI (such as algorithms) may give rise to a phenomenon termed "human alienation." This occurs when AI—a product of human creation—poses a threat to the evolution of human subjectivity across three dimensions: communication, cognition, and labor (Xie and Liu, 2023). Thus, they call for society at large to recognize the developmental limits of AI and advocate for creating controllable, safe, and reliable AI systems while promoting a collaborative evolution between human-machine societies and general AI (Huang and Lv, 2023).

Against this backdrop, I found that, despite the existence of numerous studies exploring the emotional interactions between humans and computers (Rosalind, 1997; Reeves and Nass, 1996; McStay, 2018; Marcos-Pablos and García-Peñalvo, 2022; Peng, 2024; Lai, 2023; Gan and Wang, 2024; Zhao and Li, 2023), as well as the ethical issues arising from the development of emotional AI (Gossett, 2023; Tretter, 2024; Nyholm and Frank, 2019; Xiao and Zhang, 2024; Yin and Liu, 2021; Zhang, 2024), so far, there has been no research that approached the issue from a theoretical and speculative standpoint, focusing on the definition of the dynamic emotional interaction relationship between users and AI platforms with the development of emotional AI, and further explores how it impacts human interaction paradigms. Specifically, there is a lack of theoretical discussion on the profound changes in existing societal paradigms brought about by the advancement of emotional AI.

My goal is to fill this gap and to argue that, when algorithms integrate emotional intelligence, a new type of relationship—pseudo-intimacy—emerges between users and platforms, serving as a new paradigm of human interaction that coexists with face-to-face relationships in the real world. In this pseudo-intimacy relationship, on the one hand, users and platforms achieve instantaneous emotional interaction, partially satisfying the human's desire for intimacy. However, it is also restricted by the limited development of emotional AI and human irrationality, making the human social environment more full of contradictions and tension. Consequently, the advancement of emotional AI should not only focus on technological innovation and subjective human experiences but also fully consider its impact on human social interaction paradigms. Yet, if appropriate measures are taken to address these ethical risks, I argue, nothing can fundamentally stand in the way of the progress of emotional AI.

To elaborate on my thesis, I will first focus on how the pseudo-intimacy relationship emerges and develops, and provide a realistic explanation of its definition. Then, I will conduct a summary discussion on the relevant ethical risks, and express my attitude and recommendations toward the future development of emotional AI.

2 The pseudo-intimacy relationship of user-platform becomes a new paradigm for human interaction

The human-computer society could not exist without emotions assuming the role of the glue (Gan and Wang, 2024). However, although emotions in HCI have received attention from scholars of affective computing, such as Rosalind Pickard, since the end of the last century, and "para-social relationship" has been discussed in media studies for decades, yet emotions have been marginalized in the study of user-platform relationships in the field of sociology for a long time. This is mainly due to the stereotype of "emotion-rationality" dichotomy among some scholars (Yuan, 2021). In recent years, research within social robotics has made significant strides in enhancing robots' emotional capabilities to improve their capacity for empathy and social engagement with humans (Marcos-Pablos and García-Peñalvo, 2022). Sociological theorists are increasingly recognizing that, along with the algorithmic platform's anthropomorphic development (Wu et al., 2022; Zhao and Li, 2023), the most distinct boundary between HCI and interpersonal social interaction—the authenticity of the interaction object (Giles, 2002)—has been broken. The user-platform relationship has beyond the "para-social relationship" defined by HCI scholars, resulting in a "pseudo-intimacy relationship" between humans and humanlike entities. This is evident in current HCI, where users anthropomorphize and idealize computers based on their emotional intelligence, forming social relationships that are more satisfying than face-to-face ones.

2.1 The user-platform emotional relationship is thoroughly elucidated in the context of immediate interaction, and partially satisfying the human need for intimacy

Anthropocentrism posits that humans have an inherent tendency to anthropomorphize non-human entities, driven by a desire to engage and connect with society (Epley et al., 2007). In *Alone Together,* Sherry Turkle, a professor of sociology at the Massachusetts Institute of Technology (MIT), examined the psychological phenomenon whereby individuals forge intimate connections with computers. She argues that humans can develop emotional relationships with computers, even may regard them as significant others akin to family and friends (Sherry, 2014). This human-computer relationship established on this premise—particularly in the context of social media—mimics emotional bonds found among humans. However, it lacks the depth and complexity characteristic of genuine human interactions, which is somewhat constrained by the technological advancements available at that time. Since then, scholars have increasingly suggested that individuals may integrate computers into their interpersonal networks and become emotionally reliant on their presence

(Thomas and Julia, 2018; Wang, 2023; Wang et al., 2024; Gan and Wang, 2024).

With the rapid advancement of AI's emotional capabilities and the widespread adoption of intelligent algorithmic platforms, this perspective is increasingly validated. Algorithmic technologies endowed with emotional intelligence facilitate instantaneous bidirectional interactions between users and platforms within the realm of emotional communication (Ke and Song, 2021; Hong and Huang, 2024). Based on the emotional purpose of human communication, this paper characterizes it as "pseudo-intimacy relationship." In this relationship, due to the lack of non-verbal social cues in face-to-face interactions, instant emotional interactions between users and platforms mediated by affective AI may lead users to overinterpret limited information (Walther et al., 2015), thereby leading to an unhealthy development of the relationship between the two.

In terms of emotional interaction, the enhancement of algorithmic emotional intelligence not only made algorithmic platforms novel objects of human interaction but also awakened and partially satisfied the latent human need for intimacy. Some researchers have noted that this enhancement facilitates the mobilization of human emotions for immediate user-platform emotional interactions (Bie, 2023). With emotional intelligence, users display strong conscious or unconscious emotions during interactions (Nagy and Neff, 2015), and continuously motivating themselves to engage while also eliciting immediate emotional feedback from the platform, thereby accelerating the emotional flow between them.

In addition, from the point of view of the "mirror me" theory put forward by American sociologist Charles Horton Cooley, the essence of user-platform emotional interaction is an extension of human emotion projection and the construction of the ideal self in social interaction (Gan and Wang, 2024). In the dynamic interplay of human emotional projection and computer affective computing, a recursive effect akin to an "infinite mirror" emerges between the two entities (Panaite and Bogdanffy, 2019), wherein emotions are continuously iterated and refined. This process fosters the evolution of user-platform communication forms and experiences, with pseudo-intimacy becoming a defining characteristic of the user-platform relationship. This further deepens the emotional exchange between users and platforms, potentially elevating it to a cultural level and generating consensus on granting platforms the status of "interaction subjects" in society, and even envisioning a future where user-platform emotional exchanges are equalized.

However, it must be pointed out that, in contrast to the technological object essence of emotional AI, only human beings are truly emotional animals. Emotions, as a reflection of collective human intentions, are expressed through and rationalize human behavior (Swallow, 2009). As a result, regarding the evolution of user-platform relationship attributes, emotional intelligence in AI systems is an external factor, while the human need for intimacy is the initial driving force that promotes pseudo-intimacy relationship to be a new paradigm of human interaction.

2.2 User-platform emotional interactions have become more real and tangible, while the human social environment characterized by heightened contradictions and tensions

From the perspective of social relationships, before algorithmic emotional capabilities were developed, the user-platform relationship was fundamentally an HCI. Even with emotional undertones, it remains a one-sided contribution from users, who receive no emotional response from platforms and only feedback on usage and satisfaction—referred to as "user stickiness" (Periaiya and Nandukrishna, 2024). In contemporary times, with further developments in emotional AI technology, algorithmic platforms are now endowed with emotional capabilities. The anthropomorphic affective attributes within the user-platform relationship have become more pronounced in communicative contexts (Zhejiang Lab and Deloitte, 2023). This evolution has introduced a degree of warmth into these interactions, leading to the emergence and implementation of conversational and companionable AI.

However, akin to two sides of the same coin, the development of emotional intelligence in AI systems has also introduced a range of associated risks and sparked extensive discussions regarding their ethical implications within studies (McStay, 2018; Greene, 2020; Gremsl and Hödl, 2022; Gossett, 2023; Tretter, 2024). These discussions highlight the potential benefits of emotionally capable AI systems while simultaneously addressing the challenges posed by the technological uncontrollability of AI companions and human irrationality in emotional ineractions with intelligent systems (Yang and Wu, 2024; Chen and Tang, 2024). Scholars contend that as long as emotional AI technologies can influence human emotions, they possess the potential to serve as instruments of emotional deception (Bertolini and Arian, 2020). In light of these concerns, many researchers advocate for implementing protective measures across various fields such as education, healthcare, and justice to regulate AI systems capable of interpreting and responding to human emotions while preventing their irrational use (McStay, 2020; Vagisha and Harendra, 2023; Crawford, 2021).

In the context of the user-platform relationship that this article examines, the advancement of emotional AI technology has also exacerbated ethical concerns related to private data security, algorithmic bias leading to discrimination, and information cocooning (Mei, 2024; Yan et al., 2024). This is because, as the user-platform relationship becomes increasingly emotional, the relational attributes between a given platform and its different users may differ significantly. For platforms to sustain stable user-platform relationships, they must collect extensive data on users' emotional preferences and privacy information (Lu et al., 2022). However, current legislative frameworks regarding data protection in several countries with advanced platform technology development—such as China and the United States—remain incomplete. There are no uniform norms or standards governing how interest groups backing algorithmic platforms can protect or utilize such data. Grounded in media literacy, this situation has prompted a degree of self-reflexivity among users, leading them to develop concerns about risks associated with self-information security and emotional manipulation—commonly referred to as algorithmic anxiety (Cha et al., 2022).

In addition, from the perspective of the overall social environment, the current user-platform relationship can be characterized as a pseudo-intimacy relationship that does not exist in a seamless enclosure devoid or isolated space solely created by algorithmic platforms, AI, and other emotional agents. Instead, it coexists with genuine interpersonal socialization within real society, collectively forming a social environment rife with contradictions and tensions for individuals. Therefore, while the user-platform pseudo-intimacy relationship may enrich an individual's social life and alleviate loneliness to some extent (Yuan et al., 2024), it also impacts users' real-life interpersonal relationships. This influence can even adversely affect their social skills and attitudes, thereby hindering their understanding of interpersonal emotions and their significance, reducing opportunities for establishing more meaningful interactions (Sharkey and Sharkey, 2011; Nyholm and Frank, 2019). This negative impact arises because algorithmic platforms despite being programmed to understand and react to human emotions, it still lack the same innate capacity for empathy inherent in human beings (Morgante et al., 2024). Furthermore, the natural divide between humans and computers also leads users to perceive algorithmic platforms as a "quasi-other" (Mu and Wu, 2024). In this context, true reciprocal emotional communication between users and platforms has yet to be realized. The equalization of emotional communication between the two will continue to be a lengthy and challenging endeavor, hindered by both technical and ethical constraints.

3 Conclusions and future research

In summary, I argue that as emotional AI continues to develop, the user-algorithm platform relationship has shifted from a traditional HCI to a negotiating pseudo-intimacy between humans and humanlike entities. This is not only an imaginative, anthropomorphic, and social feature that emotional AI has bestowed upon HCI, but also an important supplement to human existing interaction paradigms. In addition, the emergence and development of pseudo-intimacy relationships partially satisfy human needs for intimacy in modern society; however, due to limitations about technological development and other factors, it is not entirely beneficial and raises ethical issues like privacy data security, increasing contradictions, and tension in the human social environment.

Therefore, we should agree that the advancement of emotional AI must focus not only on technological innovation but also on ethical constraints imposed by existing social norms, such as privacy security. Technological progress that violates ethical norms is always unacceptable. Also, in the face of the increasing emotional capabilities of algorithms, we should abandon the binary thinking of technology vs. humanity, rationality vs. emotion, and explore the harmonious coexistence of humanistic spirit and technological rationality in the future (Peng, 2021).

The discussion in this paper also has some limitations. Research on integrating emotional intelligence into algorithms, exploring the development of emotional functions within AI systems through methods such as human-computer experiments holds more significant practical application value. However, due to constraints related to genre, this paper primarily summarizes and examines these concepts at a theoretical level without engaging in large-scale experimental studies. Furthermore, as previously noted, the discourse surrounding ethical issues tacitly approve that we ought to allocate ethical responsibilities to AI technologies that are integrated with emotional intelligence. The reality, however, is that the questions of whether and how to continue to refine these technologies, whether and how to assign ethical responsibilities to them, and how humans should respond to the humanlike qualities of these technologies when interacting with them, are still under development and heated discussion. The answers need to be synthesized through data, theory, and other explorations by future researchers in computing science, humanities, social sciences, and other fields. Of course, it is also possible that there will not be a definite answer for a long period, which is also the charm of academic research.

Author contributions

JW: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work has been supported by the Outstanding Innovative Talents Cultivation Funded Programs 2023 of Renmin University of China (grant number 22RXW190).

Conflict of interest

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bertolini, A., and Arian, S. (2020). *Do Robots Care? Towards an Anthropocentric Framework in the Caring of Frail Individuals through Assistive Technologies*. Berlin: Walter de Gruyter Gmb.

Bickmore, T. W., and Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput. Hum. Interact*. 12, 293–327. doi: 10.1145/1067860.1067867

Bie, J. H. (2023). Platformized digital interactions: affective practices based on the availability of technology. *Young Journal*. 4, 22–25. doi: 10.15997/j.cnki.qnjz.2023.04.007

Bie, J. H., and Zeng, Y. T. (2024). Algorithmic imagination of platform participation and affective networks: an analysis of users on Xiaohongshu. *China Youth Res*. 2, 15–23. doi: 10.19633/j.cnki.11-2579/d.2024.0018

Cha, D. L., Jiang, Z. H., and Cao, G. H. (2022). A study of user-perceived algorithmic anxiety and its structural dimensions in information systems. *Intell. Sci*. 6, 66–73. doi: 10.13833/j.issn.1007-7634.2022.06.009

Chen, S. H., and Tang, L. (2024). Human-machine love: emotional interchain and emotional intelligence coupling of artificial intelligence partners. *J. Hainan Univ*. 9, 1–9. doi: 10.15886/j.cnki.hnus.202405.0437

Crawford, K. (2021). Time to regulate AI that interprets human emotions. *Nature* 592:7853. doi: 10.1038/d41586-021-00868-5

Epley, N., Waytz, A., and Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychol. Rev*. 114:864. doi: 10.1037/0033-295X.114.4.864

Gan, L. H., and Wang, H. (2024). From emotional projection to digital emotion: emotional transformation of human-computer interaction in digital landscapes. *Mod. Publish*. 3, 27–38. Available at: https://kns.cnki.net/kcms2/article/abstract?v=64ENavj7QCDmCc1OEm2hxvh6X0oFChoroZAEOKw16Ydf4Nuh6PvnuaXoedaO7_9n5dh1kWt2tD2u33p4GFjpUaVrOcWpal7QLsmPovVfdlbAKpDm7zdIF21SETl2I9Q5m19sL9JVYZUBVCNu6i2QYQbPrEonkZ3nFAtAkaZeP_7aOJQHx_nxFduevAu-eKQVczw&uniplatform=NZKPT&language=CHS

Giles, D. C. (2002). Parasocial interaction: a review of the literature and a model for future research. *Media Psychol*. 4, 279–305. doi: 10.1207/S1532785XMEP0403_04

Gossett, S. (2023). *Emotion AI: 3 Experts on the Possibilities and Risks*. Available at: https://builtin.com/artificial-intelligence/emotion-ai (accessed September 27, 2024).

Greene, G. (2020). *The Ethics of AI and Emotional Intelligence*. Available at: https://partnershiponai.org/paper/the-ethics-of-ai-and-emotional-intelligence/ (accessed September 27, 2024).

Gremsl, T., and Hödl, E. (2022). Emotional AI: legal and ethical challenges. *Inf. Polity* 27, 163–174. doi: 10.3233/IP-211529

Hong, J. W., and Huang, Y. (2024). "Making" emotions: the logic of human-computer emotion generation and the dilemma of invisibility. *Journal. Univ*. 1, 61–121. doi: 10.20050/j.cnki.xwdx.2024.01.008

Huang, R., and Lv, S. B. (2023). ChatGPT: ontology, impact and trends. *Contempor. Commun*. 2, 33–44. Available at: https://kns.cnki.net/kcms2/article/abstract?v=64ENavj7QCDOIGKAp9OFuZLyR4yvYC43G7dFd2r6wwu6GUBFvd-uUu57d8SGfpBJDSDzCbINfbIIFqu04MVonzb_e-lHoSCyQTIgc9AypbD2ZtoSitaKGAQxil3NrzlT3p8mdMg4UWN9aC1xbVp7ssP5y2NqpalKaug-wVL1KQDKVMTGA3nS-0jjoAGO7lLhcfk&uniplatform=NZKPT&language=CHS

Ke, Z., and Song, X. K. (2021). From "me in the mirror" to "me in the fog": the aberration and theoretical crisis of social interaction in virtual reality. *Journal. Writ*. 8, 75–83. doi: 10.20050/j.cnki.xwdx.2023.02.008

Lai, C. Y. (2023). Recursive negotiation: the symbiosis and interaction between users and algorithms on short video platforms: an ethnographic study centered on "influencers". *Journalist* 10, 3–15. doi: 10.16057/j.cnki.31-1171/g2.2023.10.002

Lu, Y. Y., Sun, Y. T., Zhang, Y., and Li, X. G. (2022). Impact of artificial intelligence, machine learning, automation and robotics on the information industry—overview and implications of CILIP symposium 2021. *Libr. Intell*. 66, 143–152. doi: 10.13266/j.issn.0252-3116.2022.19.014

Marcos-Pablos, S., and García-Peñalvo, F. J. (2022). *Emotional Intelligence in Robotics: a Scoping Review*. Cham: Springer.

Marvin, M. (2006). *The Emotion Machine*. Hangzhou: Zhejiang People's Publishing.

Marx and Engels (2009). *Collected Works of Marx and Engels, Volume 4*. Beijing: People's Publishing House.

McStay, A. (2018). *Emotional AI: the Rise of Empathic Media*. London, Thousand Oaks, CA: Sage.

McStay, A. (2020). Emotional AI and EdTech: serving the public good? *Learn. Media Technol*. 45, 270–283. doi: 10.1080/17439884.2020.1686016

Mei, A. (2024). Innovation of algorithmic discrimination governance model under positive ethics. *Polit. Law* 2, 113–126. doi: 10.15984/j.cnki.1005-9512.2024.02.007

Morgante, E., Susinna, C., Culicetto, L., Quartarone, A., and Lo, B. V. (2024). Is it possible for people to develop a sense of empathy toward humanoid robots and establish meaningful relationships with them? *Front. Psychol*. 15:1391832. doi: 10.3389/fpsyg.2024.1391832

Mu, Y., and Wu, Y. H. (2024). From quasi-social relation to human-machine relation: a two-dimensional classification model based on authenticity and interactivity. *Contempor. Commun*. 3, 9–14. Available at: https://kns.cnki.net/kcms2/article/abstract?v=64ENavj7QCBTNAmv8qsiWRsZaefY95bL9wUDOq7Tvt0fUkC7sujULEcQa-N0yxrQ2HqNNyw8k7SqhCkmpZm_SFQDl_BTLArm74g8pr_azVZ6orv7M9tzXFRWMTmSSrt1GEbNKX6tFAMm09O3a3mdKN09m-jpHth_dxNfE80dnBlT7o1w_pgQhqnRw43GF0Vo&uniplatform=NZKPT&language=CHS

Nagy, P., and Neff, G. (2015). Imagined affordance: reconstructing a keyword for communication theory. *Soc. Media Soc*. 1, 1–9. doi: 10.1177/205630511560 3385

Nyholm, S., and Frank, L. E. (2019). It loves me, it loves me not: is it morally problematic to design sex robots that appear to love their owners? *Techne Res. Philos. Technol*. 23:122110. doi: 10.5840/techne2019122110

Panaite, A. F., and Bogdanffy, L. (2019). Reimagining vision with infinity mirrors. *MATEC Web Conf*. 290:e01011. doi: 10.1051/matecconf/201929001011

Paul, L. (2017). *Replaying the Human Journey: Media Evolution*. Chongqing: Southwest Normal University Press.

Peng, L. (2021). Survival, cognition, relationships: how algorithms will change us. *Journalism* 3, 45–53. doi: 10.15897/j.cnki.cn51-1046/g2.2021.03.002

Peng, L. (2024). "Mirror" and "Other": an examination of the relationship between intelligent machines and humans. *Journal. Univ*. 3, 18–118. doi: 10.20050/j.cnki.xwdx.2024.03.001

Periaiya, S., and Nandukrishna, A. T. (2024). What drives user stickiness and satisfaction in OTT video streaming platforms? a mixed-method exploration. *Int. J. Hum. Comput. Interact*. 40, 2326–2342. doi: 10.1080/10447318.2022.2160224

Reeves, B., and Nass, C. I. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge: Cambridge University Press.

Rosalind, P. (1997). *Affective Computing*. Boston, MA: The MIT Press.

Rubin, A. M., Perse, E. M., and Powell, R. A. (1985). Loneliness, parasocial interaction, and local television news viewing. *Hum. Commun. Res*. 12, 155–180. doi: 10.1111/j.1468-2958.1985.tb00071.x

Sharkey, A., and Sharkey, N. (2011). Children, the elderly, and interactive robots. *IEEE Robot. Automat. Mag*. 18:940151. doi: 10.1109/MRA.2010.940151

Sherry, T. (2014). *Group Loneliness*. Hangzhou: Zhejiang People's Publishing House.

Swallow (2009). *Emotional Culture in Chinese history: An Interdisciplinary Textual Study of Ming and Qing Texts*. Beijing: The Commercial Press.

Thomas, H. D., and Julia, K. (2018). *Human-Machine Symbiosis*. Hangzhou: Zhejiang People's Publishing House.

Tretter, M. (2024). Equipping AI-decision-support-systems with emotional capabilities? Ethical perspectives. *Front. Artif. Intell*. 7:1398395. doi: 10.3389/frai.2024.1398395

Vagisha, S., and Harendra, K. (2023). Emotional intelligence in the era of artificial intelligence for medical professionals. *Int. J. Med. Grad*. 2:112. doi: 10.56570/jimgs.v2i2.112

Walther, J. B., Van Der Heide, B., Ramirez, A., Burgoon, J. K., and Peña, J. (2015). Interpersonal and hyperpersonal dimensions of computer-mediated communication. *Handb. Psychol. Commun. Technol*. 1:22. doi: 10.1002/9781118426456.ch1

Wang, H., Hu, F. Z., Liu, I T., Gan, L. H., and Liu, T. (2024). Has the digital landscape beautified our lives? (Academic Dialogue). *Res. Cult. Art* 1, 56–114. Available at: https://kns.cnki.net/kcms2/article/abstract?v=64ENavj7QCBMlpFdshFmLBSZaXN8v3vovBTQphAfVf-EnfMRlTHUerJ3nVonChNvSvmzmM9oiiGe29RJPYhh0u7GxPsHtIzMAZUZHxlDEy3ZR7eguR6l4EJQ-3VPhn3OGjFSwlg-rg-Yfw7W86bJO6OHiWpvUsuFiQ2_5ePl1ufu2D0Got2m0BXM94TJGXIA&uniplatform=NZKPT&language=CHS

Wang, Z. W. (2023). Marx's three metaphors on machines—a study based on the perspective of human-machine relationship. *Monthly J. Theory* 9:19. doi: 10.14180/j.cnki.1004-0544.2023.09.002

Wu, J. W., Yang, P. C., and Ding, Y. H. (2022). The questioning of technology: an examination of the human-technology relationship in smart news production. *Journal. Writ*. 10, 29–42. Available at: https://kns.cnki.net/kcms2/article/abstract?v=64ENavj7QCCCyKoUtw7W4wVgMSKHUm5n17Mhzm0_xOf82fri1SP9WR8isOGCNorKMiZJdngUpJo_M2K3aLin4d4LWxTuRajI2seiugCrbxgW8nTgkz-k4glU-S7o-s0wOPBuL8vjzdmNhl81ICDZwVD_DycWRJzLz1bzoFyc_eCkCO9yMur7iF2Zit-HlmlB&uniplatform=NZKPT&language=CHS

Xiao, H. J., and Zhang, L. L. (2024). Theoretical deconstruction and governance innovation of big model ethical misconduct. *Res. Fin. Iss.* 5, 15–32. doi: 10.19654/j.cnki.cjwtyj.2024.05.002

Xie, J., and Liu, R. L. (2023). ChatGPT: generative artificial intelligence causes human alienation crisis and its rethinking. *J. Chongqing Univ*. 5, 111–124. Available at: https://kns.cnki.net/kcms2/article/abstract?v=64ENavj7QCBfxrU5noED1Y8e1Q7L3BP_ozGl6n6FWryHTMp-LJIXOELosG08rwmhn7e3JGH41P1_a8wkqKOx-5ZxsfKA0lBoj9g3V5pzDrQHeelJo4-eWk7_QJOgycM0SmrgOCfCOvTp7J5g-t3gFmLiaLj1NvZ3na3d5lN06GZrQ6lUALQnTqMnx5SB8rPf&uniplatform=NZKPT&language=CHS

Yan, W. W., Wang, Y. Y., and Song, J. H. (2024). A study on the impact of user data collection on the willingness to abandon the use of artificial intelligence services - based on information sensitivity and privacy perspectives. *Intell. Sci*. 5, 1–21. Available at: https://kns.cnki.net/kcms2/article/abstract?v=64ENavj7QCD4p611K7fU3WTCrRqcvYN3tUaVFIdG8jKJhm6z8kFVkn4U3WKXsHDVrq3CT2eUyttUSEvAx6wQopLsAV_xW7uhhkm9NPWb2ThIp4Eh8CZ7xBxkKI9PScb-tTPAI7SxlKVDvhmTAYDLI4huuETb2wy00EvxdpCkbaXbXmpVPKWrQ62FicBKtKWx&uniplatform=NZKPT&language=CHS

Yang, J., and Wu, N. (2024). Social problems of brain-computer interface technology and its countermeasures. *Ideol. Theoret. Front.* 1, 80–88. doi: 10.13231/j.cnki.jnip.2024.01.009

Yin, L. G., and Liu, Y. L. (2021). Technological empowerment and visible labor in short video platforms—an examination based on the political economy of communication. *Fut. Commun*. 6, 41–121. doi: 10.13628/j.cnki.zjcmxb.2021.06.010

Yin, Q., and Lin, Y. (2023). "Pigeons that drag the shift": elastic relations in platform content production labor—an exploratory study based on Beili Beili. *Journal. Commun. Res*. 12, 69–128. Available at: https://kns.cnki.net/kcms2/article/abstract?v=64ENavj7QCDJnxrwLhTYYsE0ZSBmyvMAm-M9DrJf_laMyHoMDKBKYj_cDZjIWqBYVV2BQ5cREWdOJqr6kZzfxsijpx_YFLGUot6HYy1SYKqlDn98VdBprdnYwY1NSB0SYXXfrwDVQAIVD0PNUtC3XCF1ux1KeUoj79174C2YtuCHLrnDbW63rSKG7NjTQPbO&uniplatform=NZKPT&language=CHS

Yuan, G. F. (2021). Toward a theoretical path to "practice": understanding emotional expression in public opinion. *Journal. Int*. 6, 55–72. doi: 10.13495/j.cnki.cjjc.2021.06.004

Yuan, Z., Cheng, X., and Duan, Y. (2024). Impact of media dependence: how emotional interactions between users and chat robots affect human socialization? *Front. Psychol.* 15:1388860. doi: 10.3389/fpsyg.2024.1388860

Zhang, L. H. (2024). Algorithmic security risks and their resolution strategies in the construction of digital society. *J. Northeast Norm. Univ.* 2, 134–144. doi: 10.16164/j.cnki.22-1062/c.2024.02.014

Zhao, Y., and Li, M. Q. (2023). Virtual anchor practice and human-computer emotional interaction under the trend of anthropomorphism. *Mod. Commun.* 1, 110–116. doi: 10.19997/j.cnki.xdcb.2023.01.012

Zhejiang Lab and Deloitte (2023). Research on the development and application of affective computing. *Softw. Integr. Circ.* 8:284. doi: 10.19609/j.cnki.cn10-1339/tn.2023.08.033

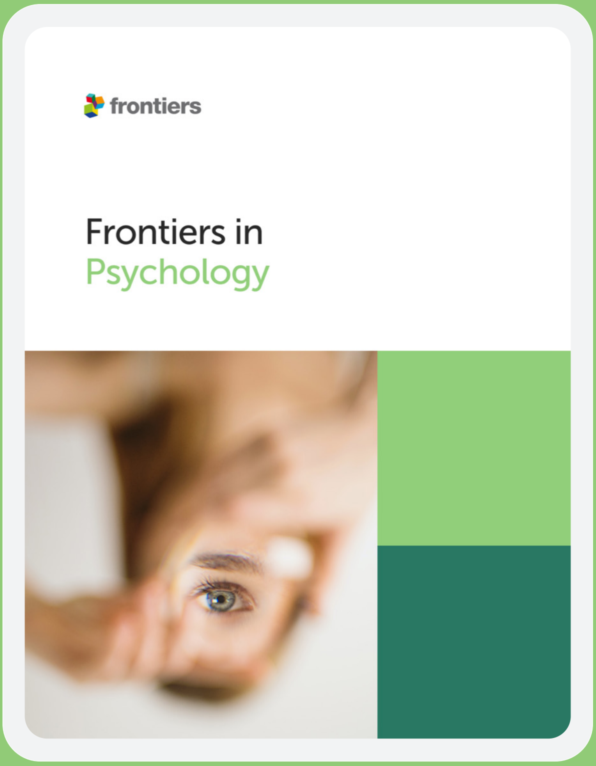
Frontiers in Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

Discover the latest Research Topics

See more →



frontiers

Frontiers in Psychology

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact