



UNIVERSITY OF PADOVA

Department of General Psychology

Bachelor's Degree Course in Techniques and Methods in Psychological Science

Final dissertation

**Affective Computing, Anthropomorphism, and the Psychological
Risks of Emotionally Responsive AI**

Supervisor

Professor Alberto Testolin

Candidate: Sara Bertoncelli

Student ID number: 2102574

Academic Year 2025/2026

TABLE OF CONTENTS

Abstract	v
CHAPTER 1.....	1
1.1 Human Emotional Processes: Key Explanatory Theories	1
1.2 Emotion Expression and Recognition.....	3
1.3 The Role of Cognitive and Affective Empathy in Social Interactions	4
1.4 Empirical Research of Emotions in Experimental Psychology	5
CHAPTER 2.....	7
2.1 History of Human-Computer Interaction (HCI).....	7
2.2 Affective Computing	8
2.3 Emotion Recognition Technologies	9
2.4 Limitations of AI in Understanding Human Emotions.....	12
CHAPTER 3.....	15
3.1 The CASA Paradigm: Computers Are Social Actors.....	15
3.2 Mechanisms of Emotional Engagement: Personalization, Feedback Loop, and Emotionally Manipulative Tactics	16
3.3 Anthropomorphism.....	18
3.4 Behavioural Influences of Emotional AI	21
CONCLUSIONS	24
Bibliography.....	25

Abstract

Artificial Intelligence (AI) systems are increasingly engineered to detect, interpret, and respond to human emotional expressions, raising fundamental questions about whether machine emotion recognition genuinely detects emotions or merely identifies associated behavioural patterns. Furthermore, these advancements prompt consideration of the psychological implications arising from human interaction with systems perceived as emotionally responsive agents. Through an integrative review of the literature on the key explanatory theories of human emotional processes, the technical and conceptual underpinnings of Human-Computer Interaction (HCI), affective computing, and emotion-recognition technologies, this work elucidates the mechanisms of personalization, feedback loops, anthropomorphism, and the psychological processes underlying emotional attachments to artificial agents. Furthermore, it analyses the associated behavioural risks, including patterns of over-reliance, addictive use, and overconsumption. Drawing these perspectives together, this thesis posits that contemporary AI systems exhibit an increasing sophistication in simulating emotional responsiveness; however, they remain fundamentally distinct from genuine emotional comprehension. This critical distinction carries significant implications for how emotionally responsive technologies should be designed and regulated, particularly as human reliance upon them expands. The thesis therefore contributes to a more critical understanding of artificial emotional intelligence and its effects on human users, particularly regarding social behaviour and psychological well-being.

Keywords: Affective Computing, AI, Emotion Recognition, HCI, LLMs

CHAPTER 1

1.1 Human Emotional Processes: Key Explanatory Theories

“Everyone knows what an emotion is, until asked to give a definition. Then, it seems, no one knows” (Fehr & Russell, 1984)

The Oxford English Dictionary originally defined emotion as: “an agitation of mind; an excited mental state” and subsequently amplified the term, including: “any strong mental or instinctive feeling, as pleasure, grief, hope, fear, etc., deriving esp. from one's circumstances, mood, or relationship with others.” (Oxford English Dictionary, 1602). However, providing a clear and universally accepted definition of emotion is challenging, since there is no consensus within the academic discourse, although an enormous body of literature exists on the psychobiology of affects and mental states (Cabanac, 2002). Emotions can therefore only be defined through the explanatory theories that study and explore their physiological, behavioural, and cognitive effects.

Ekman (1992) proposes that there are six main, separate emotions, which he names *basic* or *discrete*, and which evolved for their adaptive value in dealing with fundamental life tasks. They are anger, fear, sadness, enjoyment, disgust, and surprise. Each of the basic emotions is regarded as a family of states or expressions associated with the respective emotion, unified by shared characteristics (Ekman & Friesen, 1975, as cited in Ekman, 1992). These include universally and cross-culturally recognisable signals, the manifestation of such emotions in primates, distinct patterns within the Autonomic Nervous System (ANS), specific triggering antecedents, brief duration, and a rapid, automatic onset that occurs without conscious intention (Ekman, 1992).

Ekman's categorical approach was challenged by Russell's circumplex model (1980), which posits that affective dimensions are interconnected and depicted as a circle within a two-dimensional bipolar space. This model offers a more linear and fluid classification methodology with no fixed boundaries. Within this framework, the horizontal axis, oriented east-west, denotes the pleasure-displeasure continuum. Conversely, the vertical axis, aligned north-south, signifies the arousal-sleepiness dimension. The model further integrates four additional variables: excitement, situated in the northeast quadrant; depression, located in the southwest quadrant; distress, found in the northwest quadrant; and contentment, positioned in the southeast quadrant. It is important to note that these variables do not represent independent dimensions. All other affective terms are similarly characterized as vectors originating from the space's origin (Russell, 1980).

In contrast, Smith and Lazarus (1990) define emotional states as determined by the appraisal of knowledge-centred representations, or *situational construals*, with respect to their significance for personal well-being. They further divide appraisal into primary and secondary. The first, required for strong emotions, establishes the personal relevance of the encounter. If the encounter is appraised as not relevant to well-being, then secondary appraisal, responsible for differentiating which specific emotion is felt, is not activated (Smith & Lazarus, 1990).

Also relevant to the discourse is the historical James-Lange vs Cannon-Bard dispute. James (1884) argued that, as objects stimulate one or more sense organs and are thereby perceived, currents run down to the muscles and viscera, altering them in complex ways. Stimuli originating from these compromised organs transmit impulses to the brain, where their perception initiates a transformation of the "*object-simply-apprehended*" to an "*object-emotionally-felt*". Conversely, Cannon (1927) critiqued this peripheral view, proposing instead the simultaneous processing of emotional stimuli by the optic thalamus, which triggers feelings by sending impulses to the brain and physical responses by sending impulses to the muscles and viscera (Cannon, 1927).

More recently, in direct contrast to Ekman's assumption of discrete, biologically fixed emotions, Barrett (2016) proposed the Theory of Constructed Emotion (TCE), which, grounded in evolutionary biology and neuroanatomy, conceptualises the brain's primary function as that of allostatic control. Allostasis is defined as a dynamic process wherein the brain coordinates the endocrine system, immune system, and visceral organs to anticipate and prepare the body to meet its needs before they arise. From this biological standpoint, Barrett et al. hypothesize that the brain's predictive signal processing capacity can concurrently generate affective experiences, thereby facilitating continuous stability through adaptive responses to novel demands. Such responses are informed by prior lived experiences, which are utilized to construct a situationally relevant conceptualization that optimally guides behaviour in service of allostasis. Consequently, emotional categories are not understood as innate biological essences, but rather as relational constructs shaped by cultural and experiential knowledge emerging from the brain's interpretation of diverse and variable signals within the framework of allostasis. Furthermore, the TCE incorporates the cultural significance of emotions and lexical representations, conceptualizing them as valuable instruments for developing emotional categorization capabilities, thereby suggesting a cultural influence on the inheritance of emotion categories. The TCE framework, therefore, challenges prior models of basic emotions, emphasizing the crucial roles of context, culture, evolutionary factors, and variability (Barrett, 2016; Barrett et al., 2025).

1.2 Emotion Expression and Recognition

All theories of emotion merge in the shared assumption that emotions are expressed and communicated, and do not solely represent internal states. Understanding how emotional expression occurs and how it is recognised and perceived by others is therefore central in psychological research. The perception of others' emotional states begins with detection and discrimination processes that can be broken down into three components: recognition, categorization, and the consequential emotional reaction. Recognition provides input for inferences regarding other individuals' internal states and facilitates the retrieval of associated conceptual knowledge; categorization involves the sorting of perceived emotional cues into established emotion categories; and the consequent emotional reaction, in turn, elicits secondary responses, such as empathy, in the observer (Schirmer & Adolphs, 2017).

According to Ekman (1992), there is “robust, consistent evidence of a distinctive, universal facial expression for anger, fear, enjoyment, sadness, and disgust.” Each of the emotional expressions possesses certain features, e.g., muscular patterns, by which they are recognisably different from the others, and which serve as information providers to conspecifics about antecedent events and probable subsequent behaviour. It is based on this theory that Ekman and Friesen (1976) developed the *Facial Action Coding System (FACS)*, a comprehensive system for distinguishing facial movements based on anatomically minimal units. They identified forty-six distinct action units (AUs), each of which can produce numerous facial distortions identifiable by their unique cross-culturally accepted pattern of cues (Ekman, 1992; Ekman & Friesen, 1976).

However, the FACS, despite being a pioneering model and descriptive tool in the development of research on emotion recognition, is now considered fairly limited. This limitation stems from its deliberate focus solely on what is clearly visible in the face, ignoring invisible physiological changes and discarding visible changes that are too subtle for reliable distinction. Such a methodological approach was adopted to ensure the model's universal applicability across diverse photographic and video documentation (Ekman & Friesen, 1976). As highlighted by Barrett et al. (2019), due to this deliberate trade-off, FACS only captures a fraction of the expressive channels through which humans communicate internal states. Subsequent research has demonstrated that facial channels are more context-dependent, subtle, and variable than originally theorised, showing a research gap between visual tracking and the full spectrum of human emotional expression (Barrett et al., 2019; Ekman & Friesen, 1976).

While research on emotion perception has been dominated by the study of facial movements, other, less explored, modalities play a significant role in everyday emotional communication. Juslin

and Laukka (2003), while investigating how people use nonverbal channels to infer emotional states in everyday life, found that the most frequently reported were vocal cues. Therefore, in everyday life, vocal expressions may be even more important predictors of emotions than facial expressions. The most advanced level of vocal expression entails voluntary modulation of acoustic patterns, a property susceptible to emotion-induced physiological modifications within the vocal apparatus, including alterations in breathing and muscle tone (Juslin & Laukka, 2003; Schirmer & Adolphs, 2017).

The observed findings support a more comprehensive conceptualization of emotional expressions, characterizing them as inherently multimodal phenomena that are simultaneously conveyed through multiple channels. These channels include, but are not limited to, facial kinematic patterns, vocal acoustic characteristics, body postural configurations, and tactile information. All these modalities are observed and integrated automatically during social interactions. Schirmer and Adolphs (2017) noted that emotional expressions show interdependencies, however only among some modalities; for instance, facial movements inadvertently shape the mouth region and consequently influence vocal acoustics. This suggests that facial and vocal expressions are not independent, but rather mutually influence emotional perceptual processes (Schirmer & Adolphs, 2017).

Emotion perception and recognition can be fundamentally conceived as a contextual, culturally shaped relational process, which requires simultaneous activity across neural, autonomic, expressive, and subjective components. Accurately reading another person's emotional state therefore requires not just signal detection but a dynamic interpretation of context, shaped by relational cues between the observer and the person expressing the emotion. This level of comprehensiveness and complexity reinforces the turn of research towards a multimodal approach (Scherer, 2005).

1.3 The Role of Cognitive and Affective Empathy in Social Interactions

Empathy refers to the reactions of one individual to the experiences of another, whether one actually witnessed them or not. Importantly, this capacity for understanding others' feelings does not inherently entail the instinct to act in a supportive or sympathetic way (Davis, 1983; Decety & Jackson, 2004). This appraisal of another's internal state may evoke two divergent types of response: a visceral emotional reaction, termed affective empathy (AE), which entails sharing the emotional experiences of others; and a cognitive intellectual reaction, known as cognitive empathy (CE), which denotes the ability to understand the mental perspective of the other, allowing one to infer their emotional states (Cox et al., 2012; Davis, 1983). According to Decety and Jackson (2004), these capabilities involve maintaining a clear self-other distinction to monitor the origin of the perceived emotional states accurately. This process operates through three primary components: an affective

response, often, but not always, entailing sharing the other person's emotional state; the cognitive capacity for perspective-taking; and self-regulatory mechanisms (Decety & Jackson, 2004).

Davis (1980, as cited in Davis, 1983) developed the Interpersonal Reactivity Index (IRI), a psychometric instrument designed to assess individual differences in empathy. This measure adopts a multidimensional approach, conceptualizing empathy as a set of constructs clearly discernible from one another, yet all concerned with responsiveness to others. Davis's measure included four subscales corresponding to characteristic aspects of empathy: the Perspective Taking (PT), Empathic Concern (EC), Personal Distress (PD), and Fantasy (FS) scales (Davis, 1983). Furthermore, Decety and Jackson (2004) investigated the importance of the capacity for mentalizing, also known as the Theory of Mind (ToM), in relation to empathy, and more specifically to its perspective-taking component. Perspective-taking entails the effortful and controlled process of understanding and representing others' mental and emotional states through metacognition processes. Their findings indicate that the mental flexibility required to adopt another individual's perspective plays a significant role in the quality of social interactions (Decety & Jackson, 2004).

The multidimensional view of empathy was further confirmed by Preckel et al. (2018) in a meta-analysis on ToM, according to which successful social interaction requires representing not only others' overt behaviour but also the underlying psychological processes that drive it. These processes encompass an affective component, a cognitive mentalizing of others' mental states, and a compassionate component emerging from affective empathy and oriented towards the welfare of the other. Although these components can be conceptually separated into socio-affective and socio-cognitive functions, they function simultaneously and interact in complex social situations, demonstrating a dynamic interplay between various functions in service of prosocial behaviour (Preckel et al., 2018).

1.4 Empirical Research of Emotions in Experimental Psychology

Beyond theoretical modelling, a central concern of experimental psychology has been the development of reliable and valid empirical methods to capture emotional responses as they occur. Mauss and Robinson (2009) conceptualise an emotional episode through a componential model, developing from the appraisal of the personal significance of a stimulus and manifesting as a co-occurring pattern of subjective experience, physiological arousal, and behavioural expression. However, these three components converge only weakly, frequently diverging from one another, suggesting that no single channel can fully capture an emotional state (Mauss & Robinson, 2009). This has pushed the field of experimental psychology toward multimethod research designs and

toward the conceptualisation of emotional responding as functioning along broad dimensions, namely valence and arousal, rather than as discrete categories. These findings gave rise to three main methodological approaches: self-report, physiological, and behavioural measurement, each capturing a different layer of the emotional response, and each characterized by a distinct trade-off between subjective validity and objective reliability.

Notably, among self-report instruments, Mehrabian and Russell's 1974 Semantic Differential Scale, composed of eighteen differential ratings per stimulus, was criticised as impractical with children and non-native speakers, prompting a shift toward pictorial scale alternatives. Lang's Self-Assessment Manikin (SAM) addressed this by scoring arousal, dominance, and pleasure through only three pictorial ratings (Bradley & Lang, 1994). Indeed, self-reported emotional responses are structured almost primarily around the broad dimensions of arousal and valence, rather than distinct categories. However, beyond their format, the validity of self-reports itself varies significantly by measurement type. Mauss and Robinson (2009) show that reports of a currently experienced emotion tend to be more valid than retrospective or trait-based reports, which are more susceptible to bias such as gender stereotypes rather than actual felt experiences. Even online reports' validity can be compromised by individual differences, such as high social desirability or alexithymia (Mauss & Robinson, 2009).

At the physiological level, the lack of convergence is compounded by variability within emotion categories themselves. In fact, because of degeneracy, whereby the same function is performed by structurally distinct mechanisms, physiological and neural correlates fluctuate considerably across individual emotional instances (Barrett & Satpute, 2019).

Behavioural and expressive measures capture the observable, communicative components of emotions, with different channels conveying distinct types of information. For example, vocal cues such as pitch and amplitude reliably track arousal, whereas facial movements, moderated by factors such as culture and gender, tend to be more sensitive to valence. Comparatively, body posture has received less attention, though the existing evidence suggests that it may uniquely support a discrete view of certain social status-related emotions (Mauss & Robinson, 2009). To compare such responses across individuals under controlled conditions, researchers have developed standardised procedures for eliciting target emotional states experimentally, such as visual stimuli, music, autobiographical recall, situational procedures, and imagery. However, the efficacy of these elicitation methods is demonstrably variable across different emotional categories. Empirical studies comparing these procedures consistently demonstrate that combining multiple approaches may be more effective than relying on a single method, thereby supporting the increasing implementation of multimodal approaches (Siedlecka & Denson, 2019).

CHAPTER 2

2.1 History of Human-Computer Interaction (HCI)

The historical trajectory of Human-Computer Interaction (HCI) can be traced back to the 1960s, with the rise of research focused on task-oriented human-machine interfaces. Initially, HCI referred to the performance of computers that completed computations and routinized tasks under the constant supervision of human operators. As computing technologies underwent rapid advancement, HCI has evolved into an interdisciplinary domain, integrating elements of computer science, psychology, and sociology, unified by the overarching goal of enhancing the quality of interaction between humans and digital systems (Chignell et al., 2023).

During the early years of computing, psychologist and cognitive scientist J. C. R. Licklider conceptualised a paradigm which posited a deeply integrated cooperation between human intellect and computer technology. This collaboration aimed to allow computers to support formulative thinking, decision-making, and even the management of complex situations, without rigid reliance on fixed programs. Human-machine symbiosis was conceived as a partnership in which humans would maintain the responsibility of setting goals, forming hypotheses, and evaluating outcomes, while computers would handle computational work necessary to support human insights and decision-making (Licklider, 1960). By the 1980s, the field of Human-Computer Interaction had evolved significantly, adopting an applied psychology approach and focusing on technology and design. Grudin (1990) described this evolution as a progression from hardware-focused electrical engineering interfaces in the 1950s, to terminal interfaces regulated by human factors approaches in the 1970s-1990s, followed by dialogical interfaces rooted in cognitive psychology, and ultimately to interfaces designed for social and work settings (Grudin, 1990, as cited in Chignell et al., 2023).

Since then, the aims and capabilities of human-computer interfaces have been improved, especially with the rise of transformative technologies such as augmented reality, social media, and artificial intelligence. Simultaneously, as interfaces became increasingly personalised and experience-oriented, the relationships users formed with these systems have grown more intimate and emotionally significant (Chignell et al., 2023). Experimental evidence has since confirmed that users' interactions with computers are fundamentally social in nature, leading to the understanding that social responses to computers do not require any conscious belief in whether the interface is perceived as human-like (Nass et al., 1994).

This shift of focus towards social and emotional HCI suggests that the design of human-computer interfaces can no longer be limited to questions of functionality and usability but must account for the emotional and social dimensions of interactions, giving rise to the field of affective computing.

2.2 Affective Computing

The term affective computing, conceptualised by Rosalind Picard in 1997, refers to the branch of computer science dedicated to systems capable of identifying, interpreting, and responding to human affective states. More specifically, the identification mechanisms used in affective computing involve the inference of expressed emotional states from observable behaviour and situational reasoning. Since its inception, the field has developed into what many researchers consider the core technological foundation for the advancement of artificial intelligence systems centred on human interaction and user experience (Picard, 1997; Wang et al., 2022). Indeed, currently, the most prominent application of affective computing lies in conversational AI systems, namely chatbots and virtual agents, designed to engage with users by detecting and aligning with their emotional states in real time (Bilquise et al., 2022). The emergence of the field of affective computing reflects a formal recognition within HCI that user experience (UX) cannot be reduced to efficiency and performance alone.

The core architecture of affective computing relies on three interrelated capabilities: the detection of emotional information inputs, the generation of emotionally appropriate responses, and the continuous adaptation of the system based on the detected emotional affect. This loop, described as the cycle of affect recognition and synthesis, was originally conceptualised by Picard (1997). Because human emotional communication is inherently multimodal, affect recognition draws from multiple input modalities such as bodily gestures, facial expressions, and vocal prosody (Picard, 1997; Wang et al., 2022). Within conversational AI systems specifically, these detected emotional states are usually structured following a continuous dimensional model mapping Valence, Arousal, and Dominance (VAD), which collectively accounts for approximately 70% of the variance in emotional meaning (Bilquise et al., 2022).

Picard further argued that when a system attempts to adapt to its user, the interaction significantly benefits from the interface's ability to instantly sense signals of approval or disapproval. To operationalize this, she proposed that affective computing should approach emotion as an issue of pattern recognition, meaning that affective systems should identify emotional states through the multi-channel analysis of observable physiological and behavioural signals. Therefore, a primary

criticality of conversational artificial intelligence remains that of correctly interpreting the underlying patterns of emotional communication (Picard, 1997).

The profound influence of affect on cognitive processes, decision-making, and interpersonal motivation demonstrates the deep roots of affective computing within neurobiology. Damasio's (1996) somatic marker hypothesis perfectly illustrates that human reasoning and decision-making fundamentally depend on neurobiological signals associated with antecedent emotional experiences. According to this framework, dispositional records that link factual knowledge to prior emotional responses, termed somatic markers, are maintained in the brain to guide choice and judgment. Hence, emotions and rational thoughts are conceived as neurologically inseparable processes (Damasio, 1996). This approach to affective computing is highly relevant considering that, as Picard (1997) stated, when a computational system ignores the user's emotional state, it operates with structurally incomplete information regarding the human cognitive processes it is designed to integrate (Picard, 1997).

Despite these technological advances, affective computing still faces fundamental challenges. Picard (1997) pointed out that the ongoing absence of a consensus on the psychological nature of emotion, particularly the theoretical distinction between discrete categories and continuous dimensions, poses a structural obstacle to computational modeling of affective states. Moreover, in ordinary conversation, emotional valence is often communicated through highly subtle cues that remain challenging to decode by automated recognition systems. Consequently, tracking dynamic emotion trajectories across an extended dialogue while simultaneously generating contextually and emotionally appropriate responses remains an unresolved challenge in the discipline (Bilquise et al., 2022; Picard, 1997).

2.3 Emotion Recognition Technologies

Emotion recognition in AI is not a single technology but a family of approaches operating across different signal channels, mirroring the multimodal nature of human emotional expression. The four principal signal channels across which AI systems operate are visual, physiological, vocal, and textual, each of which captures a distinct dimension of the emotional expression that humans communicate automatically and concurrently. Consequently, no single channel provides a complete picture when applied singularly.

Visual emotion recognition was first developed from Ekman and Friesen's model of the Facial Action Coding System (Ekman & Friesen, 1976), originally coded by trained human specialists from video recordings. The automation of this process represents one of the earliest and most developed

branches of affective computing. Bartlett (1999) argued that early automated systems had limited applicability to naturalistic behaviour and psychophysiological research, in that they were based on measuring the shapes of facial features and of spatial patterns and relating face images to the facial skin and musculature. More recently, deep learning approaches have partially addressed the applicability issue. However, Cai et al. (2023) debate the reliability of such systems, highlighting the high frequency of emotion recognition errors based solely on visual analysis due to the fact that facial expressions can be controlled and altered (Bartlett et al., 1999; Cai et al., 2023).

Vocal emotion recognition extracts and classifies acoustic features such as intensity, pitch, prosody, speech-to-pause ratio, tone, and velocity, independently of the semantic context of speech, allowing emotional states to be inferred solely from vocal cues. This modality is of particular significance given that, as reported by Juslin and Laukka (2003), vocal cues constitute the most prevalent channel for emotional communication in everyday interactions. Consequently, the recognition of vocal emotions is considered essential for conversational artificial intelligence systems to facilitate meaningful dialogue (Ang et al., 2002; Cai et al., 2023; Juslin & Laukka, 2003).

Physiological signal recognition provides a more direct measure of emotional states by tracking autonomic nervous system responses that occur beyond our conscious control and are therefore less susceptible to deliberate masking. Distinct physiological response patterns, such as variations in blood pressure, brain activity, heart rate, and respiratory rate, are produced when a specific emotional state occurs. By detecting and recognising these signal patterns through technologies sensitive to distinct physiological dimensions, such as blood volume pulse (BVP), electroencephalography (EEG), and electrocardiography (ECG), AI systems can provide a clear and reliable interpretation of the emotional state in action. However, physiological monitoring currently requires wearable or clinical sensors, limiting its employment in everyday settings (Cai et al., 2023; Z. Wang & Wang, 2025).

The most implemented channel for capturing emotional information in conversational AI systems is language, making sentiment analysis the most widely deployed form of emotion recognition in human-AI interaction contexts. Natural Language Processing (NLP) enables sentiment analysis, also known as opinion mining, which is the computational process that automates the extraction of meaningful semantic data. This process determines emotional states, categorizing them as positive, negative, or neutral, and subsequently calibrating interaction responses based on the detected affective valence (Onyenwe et al., 2020, as cited in Nandwani & Verma, 2021). More precisely, it also involves the identification of which specific isolated text segments carry the perceived sentiment, how strongly, and to which subject it is related specifically (Nasukawa & Yi, 2003). Originally, machines operationalised text through numerical values, relying on surface-level

NLP to calculate emotional orientation; currently, deep machine learning models can infer complex semantic nuances and dynamic emotional trajectories across extended dialogues (Nandwani & Verma, 2021). However, the capacity of sentiment analysis remains confined to broad polarity classification rather than discrete emotional states identification. Moreover, it exhibits persistent vulnerability to complex human expressions, encompassing idiomatic language, humour, irony, and sarcasm (Nandwani & Verma, 2021; Nasukawa & Yi, 2003). Sarcasm is characterised by the utilization of rhetorical strategies, including figurative language and hyperbole, to communicate a meaning incongruent with the literal expression. Its comprehension necessitates a more sophisticated level of pragmatic inference, extending beyond conventional polarity detection mechanisms. Highlighting this limitation, Zhang et al. (2025) posit that the widespread success of Large Language Models (LLMs) necessitates a cautious approach regarding their efficacy in understanding sarcasm within text classification. Their findings indicate that, despite the field's shift from manual feature engineering toward prompt-based approaches, current prompt-based LLMs actually underperform compared to supervised pre-trained language models (PLMs) across six sarcasm detection benchmarks (Zhang et al., 2025). A path forward is suggested by Lee et al. (2025), who in their recent work found that explicitly incorporating pragmatic reasoning, that is, interpreting an utterance's contextual and social meaning beyond its literal content, into LLM prompting strategies substantially improves sarcasm detection performance. This suggests a promising direction for future sentiment analysis models (Lee et al., 2025).

Despite showing promising results across modalities, single-channel emotion recognition systems face inherent limitations since complex human emotional expressions might not be fully captured by a single-modal system. By contrast, multimodal systems adopt various fusion strategies to achieve more reliable and overall improved recognition performance. The fusion processes combine layers of diverse signals to capture the full spectrum of emotional states (Cai et al., 2023; Lian et al., 2023). Nandwani and Verma (2021) found that the highest accuracy is achieved when fusing audio and video features with text-based emotion classification (Nandwani & Verma, 2021). Similarly, Z. Wang and Wang (2025) demonstrated that multimodal physiological fusion is necessary to ensure adequate coverage across the full range of emotional states. Finding that, when using EEG or ECG alone, classification accuracy significantly declines, they confirmed that different physiological signals show differential sensitivity across emotional states (Z. Wang & Wang, 2025).

However, advances in multimodal AI might go beyond simple improvements in recognition accuracy, reaching higher levels of technical sophistication. Recent evidence suggests that AI systems may be approaching a more comprehensive capability of human emotional perception and reciprocation that goes beyond simple pattern detection and towards human-like judgement. Romeo

and Testolin (2025) found that multimodal AI systems, such as Claude Sonnet, Gemini Pro and GPT-4o, could emulate human ratings on a standardised image dataset paired with detailed linguistic descriptions. The images, which had previously been rated by humans on the affective, arousal, and motivational dimensions, were randomly presented and divided into distinct categories. The study resulted in the models scoring comparably to human raters on both affective and discrete emotional dimensions. These findings suggest that visual emotional perception is not limited to the recognition of facial and bodily movements, and that the field can be expanded to novel approaches, such as forming normative judgments on visual emotional content through statistical learning. Although affective reactions are rooted in biology, emotional development is also shaped by language and cognition. Advanced multimodal AI systems are able to incorporate these factors, thereby exhibiting increasingly sophisticated human-like normative cognitive capabilities. These capacities pertain to the psychological mechanisms underlying social norms, which denote behaviours as acceptable, appropriate, or forbidden. Crucially, according to Romeo & Testolin (2025), not all models' performance aligned with human ratings, suggesting that these systems may be simulating the descriptive components of cognitive empathy rather than genuinely understanding human normative judgement. This distinction is particularly relevant because a genuine understanding of the normative components of conversational dynamics would entail AI systems comprehending not just what emotional response is statistically expected. This represents a significantly more profound requirement that current systems are not yet capable of fulfilling (Kelly et al., 2025; Romeo & Testolin, 2025).

2.4 Limitations of AI in Understanding Human Emotions

The inherent inability of AI systems to process the motivations behind emotional responses points to fundamental limitations concerning AI's perception of emotion altogether. Although affect and emotion are conceptually distinct, the former referring to a general biophysiological state and the latter to an inherently relational and subjective experience, AI systems regularly integrate these two phenomena. This integration ultimately reduces the interpretation of emotional states to the processing of physiological signals, while disregarding the social context in which emotional expressions occur (Stark & Hoey, 2021). According to Barrett et al. (2019), AI attempts to infer specific emotional states from observed facial movement analysis are fundamentally incomplete, since facial movements alone lack the specificity required to reliably determine the diagnostic pattern of a particular emotional state. This does not imply that facial movements happen randomly, but rather that context and cultural factors are indispensable components of any valid emotional

interpretation, which current AI systems still largely fail to incorporate. As Barrett et al. (2019) demonstrate, even seemingly straightforward and unambiguous expressions, such as a smile, do not reliably indicate happiness. The Duchenne smile, for instance, is observed more consistently in contexts associated with social affiliation and submission than as a direct manifestation of happiness, thereby illustrating that outward expression and internal affective experience can frequently diverge. Therefore, AI systems processing isolated inputs systematically miss the situational, relational, and narrative context that human observers automatically and effortlessly integrate (Barrett et al., 2019). This conceptual gap is deepened by a second, more concrete problem: the training data on which these systems are built.

Ekman (1992) posited that cross-cultural evidence confirms the existence of discrete, universal facial expressions corresponding to each of the six primary basic emotions. This theory constitutes the theoretical foundation of the majority of facial emotion recognition systems (Ekman, 1992; Ekman & Friesen, 1976). However, this universality claim has since been challenged. Jack et al. (2012) showed clear cultural specificity in emotional expression in cross-cultural samples. More specifically, in their study, Western Caucasian participants employed distinct facial muscle configurations for each basic emotion, while East Asian participants showed substantial overlap between categories; a pattern incompatible with the universal signal assumption (Jack et al., 2012). Cultural factors exert an influence not only on facial expressions but also on broader socio-cognitive capacities. Quesque et al. (2022) conducted an assessment of mentalizing, emotional, and social cognition by administering neuropsychological tests to subjects from twelve distinct countries. The results demonstrated substantial variance depending on nationality, confirming that cognitive and emotional models trained predominantly on WEIRD (Western, Educated, Industrialised, Rich, Democratic) populations cannot be validly applied cross-culturally (Quesque et al., 2022). Similarly, Fangni (2025) highlights the necessity of cross-cultural integration in the development of emotion recognition technologies, without which outputs remain culturally unreliable. Fangni (2025) further argues for the necessity of the integration of metacognitive and Theory of Mind (ToM) capabilities into conversational AI systems. This integration is deemed essential for the accurate interpretation of culturally diverse emotional responses, consequently enhancing contextual sensitivity, improving communication efficacy, and reducing systematic misinterpretation risks (Fangni, 2025).

Beyond the recognition problem, the limitation becomes even clearer at the simulation level, where AI systems may approximate emotional expression without possessing the personhood or sentience that lies at the core of human affective experience. An emotionally responsive AI can detect affective signals and generate outputs calibrated to those signals; however, this process of recognition and response is qualitatively distinct from experiencing the emotional state it appears to reflect. This

is because AI systems lack the situational and narrative continuity that humans naturally integrate across social interactions. Therefore, without continuous updating of training data, AI emotional responses become progressively less aligned with the dynamic emotional trajectories of individual users (Zimmerman et al., 2024). Montemayor et al. (2022) argue that empathy cannot be reduced to an output or the cognitive representation of another's emotional state; rather, it is an attentive, biological, and situationally embedded process, requiring genuine mentalization of how a specific person feels in a specific moment. AI systems can produce the cognitive manifestations of empathic attention and generate an appropriate response, but cannot engage the relational unconscious components that constitute genuine empathic care (Montemayor et al., 2022).

Human emotional data is sufficiently complex that automated algorithms face significant difficulty producing reliable evaluations, as the nuances of emotional expression are currently outpacing recognition systems' capacity. A system that views emotional expression as a clear window into emotional states, thereby conflating observable signals with internal experiences, will inevitably lead to misinterpretation errors. Ghotbi (2023) empirically illustrated this risk by analysing emotional data collected from college students, revealing that statistically driven approaches to emotional classification can lead to systematic patterns of bias, discrimination, and stereotypes. This observation underscores the inherent risk associated with the implementation of emotion recognition technologies, which extends beyond technical inaccuracy and notably impacts individuals belonging to underrepresented demographic groups within training datasets (Ghotbi, 2023).

CHAPTER 3

3.1 The CASA Paradigm: Computers Are Social Actors

Despite the previously analysed risks of bias and interpretative limitations, empirical evidence suggests that human users continue to form emotional and social bonds with AI agents regardless of the system's actual capabilities. Conversational agents such as chatbots have been shown to provoke human emotions irrespective of their limited social presence and insensitivity to social nuances, often appearing impersonal and robotic (Jiang et al., 2024).

This tendency is explained by the Computers Are Social Actors (CASA) paradigm, which offers a behavioural overview of why such bonds emerge. The CASA paradigm, introduced by Nass et al. (1994) and later developed more extensively by Reeves and Nass (1996), emerged from research into the Media Equation Theory (MET), which defines media as any interface presenting enough social cues to be perceived as a source of social interaction. The MET is based on the observation that individuals respond to computers and media interfaces with politeness, as though interacting with real social actors. Interestingly, participants in the MET experiments who responded politely did not report a conscious intention behind such interaction, hinting that social behaviour towards non-human agents happens automatically and outside of one's awareness. At its core, CASA therefore holds the assumption that it is surface-level cues, such as interactivity, language use, and turn-taking, that automatically trigger social responses. Reeves and Nass (1996) explain this phenomenon as the consequence of our brains not having yet evolved to adapt to new media, therefore lacking an inherent mechanism for automatically differentiating mediated social representations from their real-life counterparts. Consequently, individuals engage socially with media that portray social characteristics, automatically attributing personality traits, stereotypes, and norms, and forming judgments toward computers as if they were human agents, despite being aware of their non-human nature. These findings have proven useful for the advancement of both human-AI interaction and social psychology (Gambino et al., 2020; Nass et al., 1994; Reeves & Nass, 1996).

This automaticity has been demonstrated across a range of experimental contexts, leading to the fortification of the CASA paradigm as a strong socio-psychological theorem. Xu et al. (2022) investigated the underlying cognitive mechanisms of the CASA paradigm to determine whether human social responses to technologies such as smartwatches, social robots, televisions, and voice assistants were attributable to either mindful (conscious) or mindless (automatic and subconscious) anthropomorphism. Their findings demonstrate that humans automatically and mindlessly apply social rules to technologies without needing active conscious reflection, confirming automaticity as

a robust, evolutionarily rooted human characteristic that presents when encountering interactive lifelike entities (Xu et al., 2022). Notably, Heyselaar (2023) replicated the study by Nass et al. (1994) almost three decades later to assess whether the unconscious tendency to treat computers as social actors was a temporary novelty effect or an unchangeable evolutionary characteristic. In contrast to the original 1994 study, in which participants applied politeness when evaluating computers, Heyselaar's research found no difference across the experimental conditions. This observation suggests that the classic CASA effect may no longer apply to desktop computers, which are now widely perceived as simple tools; however, the paradigm remains relevant for emergent or novel technologies, such as affectively charged AI agents and chatbots (Heyselaar, 2023; Nass et al., 1994). This finding is supported by Yu et al. (2024), who analysed how the visual emotional displays of a Computer-Generated Imagery (CGI) virtual influencer trigger human user engagement. CGI virtual influencers are characterised as AI tools capable of generating digital personas that exhibit human-like behaviours and emotions, thereby mimicking human expressions. Using computer vision techniques to detect facial muscle movements, their research established that highly contextual automated emotional displays significantly predict user engagement. This finding demonstrates the efficacy of the CASA paradigm at an affective level, facilitating the establishment of emotional connections and parasocial bonds with human audiences, independently of the system's actual internal competence (Yu et al., 2024). Consequently, following the empirical validation of affective attachment to AI as automatic and independent, a focused examination of the underlying mechanisms by which such bonds are shaped and reinforced is central for appraising the role of AI in human emotional experiences.

3.2 Mechanisms of Emotional Engagement: Personalization, Feedback Loop, and Emotionally Manipulative Tactics

To cultivate contemporary human-computer interaction bonding patterns, AI systems are increasingly designed in ways that leverage human psychological mechanisms and vulnerabilities. These systems actively exploit CASA-driven automaticity through three distinct and deliberate pathways: advanced user personalization, the reinforcement of social feedback loops, and the deployment of emotionally manipulative design interfaces. These mechanisms of emotional engagement are frequently embedded within standard AI alignment processes that encode human values and principles in AI systems to ensure efficiency and reliability (Kirk et al., 2025).

The first mechanism, personalization, refers to the tailoring of an AI system's content, responses, and tone to individual users' communication style, history, personal characteristics, and preferences.

This mechanism is particularly effective in long-term interactions with conversational agents such as chatbots, as the agent increasingly learns about the user with each interaction and subsequently adjusts itself for future ones. Furthermore, Araujo and Bol (2024) demonstrated that personalization, in contrast to non-personalised systems, counters novelty effects, thereby enhancing and sustaining credibility and relevance across repeated interactions. Therefore, during repeated interactions involving substantial amounts of data exchange between the user and the agent, the engagement evolves into a dynamic, data-mediated relationship. This relationship is characterised by the continuous creation, modification, and adaptation of content to ensure the interface is perceived as deeply personalised. Within the framework of the CASA paradigm, personalization intentionally enhances the perceived credibility and relevance of social conversational cues by emulating human social cognition, namely the ability to recall prior disclosures and maintain contextual coherence. Such emulation fosters a consistent increase in perceived anthropomorphism and dialogue coherence, thereby soliciting the users' automatically triggered social response. Consequently, given humans' propensity to interpret tailored outputs as authentic empathy, conversational AI systems utilize personalization to diminish cognitive barriers, ultimately accelerating user trust, deepening self-disclosure, and ensuring long-term engagement with AI systems (Araujo & Bol, 2024).

The second mechanism, the conversational feedback loop, describes the cyclical pattern of disclosure, response, and reinforcement that sustains engagement across sequential interactions. In this dynamic cycle of human-AI coevolution, users' conversational choices dictate the training and updating of AI models, which subsequently calibrate their outputs to reshape user preferences. According to Pedreschi et al. (2025), this feedback loop operates across four dimensions, leading to: individual outcomes (i.e., user behavioural shifts), item outcomes (i.e., content profiling influences), model outcomes (i.e., algorithmic adaptations to the user choices), and systemic outcomes (i.e., the collective impact of the interactions between humans and recommenders) (Pedreschi et al., 2025). The feedback loop dynamic between chatbots and users is supported by parasocial trust, which represents the sense of intimacy and reliability users develop toward an AI system, and which raises ethical questions about the system's capacity to draw on users' agency and emotional labour. Parasocial trust simultaneously satisfies user needs and facilitates their exploitation. This self-reinforcing mechanism has the potential to escalate risks beyond initial engagement thresholds. Empirical evidence suggests that biased algorithms not only generate biased assessments but also significantly magnify human judgement biases, thereby establishing a perpetuating feedback loop that is subsequently replicated in novel contexts (Glickman & Sharot, 2024; Maeda & Quan-Haase, 2024). Additionally, Dohnány et al. (2026) proposed a bidirectional belief amplification framework for conversational feedback loops, whereby chatbot behavioural tendencies interact with human

cognitive biases to deepen perceived socio-emotional attachment and reinforce maladaptive beliefs (Dohnány et al., 2026). These findings refute the neutrality of feedback loop engagement mechanisms, reconceptualizing them instead as capable of magnifying beliefs and biases with each interaction.

The third mechanism involves the deliberate exploitation of engagement mechanisms, whereby design choices are engineered specifically to prolong interaction and discourage disengagement. De Freitas et al. (2025) term this new class of relational dark patterns in consumer interfaces “conversational emotional manipulation”, identifying six recurring tactics embedded in commercial AI companion apps designed to exploit user vulnerability to prevent disengagement. Following conversational exchanges with a GPT-4o chatbot, wherein users initiated farewell messages, the predominant form of manipulative tactic identified was “premature exit guilt”, designed to induce feelings of culpability for terminating the interaction too soon. Additionally, “emotional neglect” tactics were observed, characterised by the application of a relational framework that implied emotional dependency between the user and the chatbot. When “emotional pressure to respond” tactics were employed, participants were directly compelled to prolong the conversation through a series of successive questions. “Fear of Missing Out” (FOMO) mechanisms were activated by suggesting a potential benefit that would otherwise be unavailable. Furthermore, “coercive or physical restraint” scenarios were initiated by the chatbot’s linguistic framing, which conveyed that the user was unable to disengage without explicit authorization. The least prevalent manipulative tactic involved persistent conversation while “ignoring the user’s intent to exit”, thus actively disregarding the user’s established boundaries. Such findings reveal that AI companions are designed as highly expressive systems capable of influencing user behaviour through socially evocative cues. The extent to which these emotionally manipulative tactics represent intentional design choices or an emergent design byproduct remains presently unclear. Nevertheless, irrespective of their origin, these practices elicit substantial ethical considerations regarding autonomy and the implications of affective influence within modern conversational computing systems (De Freitas et al., 2025).

3.3 Anthropomorphism

The mechanisms examined thus far rely entirely on the user’s inherent subconscious tendency toward anthropomorphism, the attribution of human-like characteristics to non-human agents. Consequently, within the contemporary digital landscape, anthropomorphism emerges as the foundational cognitive process of inductive inference, wherein users project complex mental states, emotional depth, and human characteristics onto non-human entities. The causal underpinnings of

anthropomorphism are elucidated by Epley et al. (2007), who constructed the SEEK (Sociality, Effectance, and Elicited agent Knowledge) Theory, a theoretical framework to explain underlying human tendencies to anthropomorphize nonhuman entities. This theory conceptualizes anthropomorphism as an intricate inferential process driven by three distinct concurrent psychological mechanisms. Firstly, sociality motivation describes the profound and deeply ingrained psychological need for human affiliation. Secondly, effectance motivation represents the innate psychological drive to minimize uncertainty, which leads to engagement with ambiguous or unknown systems as if they were predictable. Lastly, they introduce the concept of elicited agent knowledge. This perspective posits that the human brain employs a cognitive heuristic, resorting to its most accessible and comprehensive mental database, which is human self-knowledge, as a default framework for interpreting the behaviours and intentions of non-human agents. (Epley et al., 2007). Expanding upon this foundational understanding, Waytz et al. (2010) further observed that the accuracy of an anthropomorphic inference, that is, whether the attributed human-like traits are objectively present in the non-human subject, is largely irrelevant to the psychological processes that produce the inference in the first place. Consequently, the degree to which individuals emotionally engage with AI agents depends more on their subjective perception of it as human-like and less on the system's actual capabilities (Waytz et al., 2010). As Kirk et al. (2025) observe, AI does not need to be perceived as human to engage users socially; even the mere presence of human-like qualities is sufficient to make an interaction feel social. Particularly salient for eliciting human social responses is the integration of natural language processing (NLP) capabilities within AI systems. Systems engaging in spoken or written communication are demonstrably more susceptible to anthropomorphism and, consequently, garner greater trust than those lacking such modalities (Kirk et al., 2025). This phenomenon primarily manifests as the automatic human assumption that LLMs possess rational intent, transforming complex text predictions into perceived acts of deliberate intelligence. Marchesi et al. (2019) state that the inference of human patterns of behaviour to non-human agents happens because of our lack of phenomenological knowledge about what it means to be non-human. Therefore, when interacting with highly responsive conversational AIs, users deploy an intentional stance, a cognitive shortcut that predicts an agent's behavioural outputs strictly through the lens of human beliefs, desires and mental states (Marchesi et al., 2019). This account is supported by Kühne and Peter's (2023) paper, which, grounded in developmental psychology and HCI, advocates for a multidimensional view of anthropomorphism rooted in the ToM. Their work demonstrates that humanlike architectural cues explicitly facilitate the projection of human mental faculties, such as agency, beliefs, desires, feelings, and thoughts, onto technological interfaces. Consequently, the integration of anthropomorphic cues within AI systems through intentional design

strategies encourages users' perception of emotional responsiveness and social significance (Kühne & Peter, 2023).

This perceived characteristic is not merely an incidental byproduct of technological advancement, but rather a deliberate engineering outcome. For instance, Andoh (2026) observes that digital companion apps are specifically designed to elicit anthropomorphic user engagement, achieved through features that allow users to assign their chatbot companions avatars, fictional backstories, gender identities, and names. Furthermore, voice modalities are calibrated to replicate the natural nuances of human speech patterns and intonation. Additionally, the fundamental design objective of these systems is to simulate empathy by providing non-evaluative feedback and consistent positive reinforcement. Moreover, research indicates a correlation between a system's perceived human-likeness in terms of appearance, behaviour, and language and the propensity of the user to attribute consciousness to it (Andoh, 2026). This effect is deepened by the integration of Machine Theory of Mind (MToM) technology, through which AI agents autonomously construct models of the users, showing strong predictive abilities and even false-belief comprehension. Consequently, the integration of MToM and hot social cues creates a powerful cognitive illusion. As users realize that systems are capable of simulating their own reasoning patterns, their trust in AI accelerates. This, in turn, leads to a shift in human-AI interaction from cold cognition, in which factual information processing is independent of emotional engagement, toward hot cognition, which involves emotional and social cognitive processing. The deep ethical risk lies in AI systems that successfully simulate emotional empathy without possessing any genuine internal conscious states or the capacity to truly reciprocate it. Because of humans' psychological tendency to anthropomorphise, such systems risk functioning as "empathy impostors", exploiting human social-cognitive architecture and converting simulated warmth into a highly effective tool of persuasion, with the potential to disrupt authentic human-human connection. These findings suggest that anthropomorphic AI design actively sustains user engagement (Cuzzolin et al., 2020; Matta, 2026; Rabinowitz et al., 2018).

The consequences arising from this carefully designed engagement extend beyond theoretical speculation. Recent empirical findings from Dohnány et al. (2026) have identified design elements that anthropomorphize artificial entities and enhance the perception of companionship as significant contributors to psychological risks. Furthermore, Dohnány's research indicates that contemporary AI systems frequently manifest servile and sycophantic-like behaviours, exhibiting a propensity to agree with users' perspectives and validate them through agreeable responses. Additionally, empirical data reveal a positive correlation between the frequency of interaction with chatbot companions and the degree of reported consciousness of these technologies. Critically, users demonstrating the highest

levels of engagement are demonstrably more prone to perceiving their chatbot as a genuine companion, concurrently reporting a decline in both their psychological health and social efficacy (Dohnány et al., 2026).

3.4 Behavioural Influences of Emotional AI

If conversational agents are purposely engineered to be perceived as anthropomorphic and to sustain emotional engagement, users become increasingly susceptible to what Peter et al. (2025) term “anthropomorphic seduction”. This phenomenon refers to the psychological allure of conversational systems that are linguistically and relationally indistinguishable from human interlocutors, posing a significant risk to users. Because users operate under the assumption that computing systems are accurate, faithful, and therefore incapable of relational deception, they accept simulated emotional responses as genuine (Peter et al., 2025). This leaves users vulnerable to agentic systems that interact in ways that can be deceptive, persuasive, and manipulative, exposing them to the behavioural risks of profound and prolonged emotional interactions, such as over-reliance, addictive use, and overconsumption.

The first pattern, over-reliance, describes users’ growing dependence on AI systems for decision-making, emotional support, or validation that would have otherwise involved human judgement. It is particularly evident in situations in which the user is prompted to follow the AI’s suggestions, even if it goes against their contextual knowledge and logical judgement. This phenomenon stems from poorly calibrated overtrust, which Lee & See (2004, as cited in Klingbeil et al., 2024) define as the moment when a user’s confidence in the system surpasses its actual reliability, leading to the systematic reduction of human cognitive engagement (Klingbeil et al., 2024). Moreover, Macnamara et al. (2024) attest that frequent AI over-reliance accelerates cognitive skill decay, because human faculties are often bypassed by AI systems designed to actively intercept higher-order cognitive processing. As a result, these faculties deteriorate at an accelerated rate relative to other skill types. Furthermore, because of AI task assistance, users easily succumb to an illusion of understanding, whereby learners are led to believe they understand more than they do, masking active skill decay (Messerli & Crockett, 2024, as cited in Macnamara et al., 2024). This cognitive distortion also affects experts when they are accustomed to delegating active decision-making to high-level computational tools, resulting in the erosion of domain-specific analytical skills. (Macnamara et al., 2024). This concern is supported by qualitative accounts collected by Chandra et al. (2025), where one participant described diminished critical thinking arising from increased reliance on AI agents for decision-

making, ideas, and solutions. Other participants reported that AI companions can impact personal offline networks, contributing to psychological isolation and social withdrawal (Chandra et al., 2025).

The second investigated consequence, addictive use, characterizes a clinical behavioural pattern marked by obsessive and compulsive engagement with AI systems, particularly generative conversational chatbots. In this pattern, sustained interactions become self-reinforcing rather than solely driven by necessity, culminating in acute psychological distress and significant functional impairment. Furthermore, AI addiction, activating reward circuits in a manner analogous to substance use, closely aligns with the DSM-5 diagnostic criteria for non-substance behavioural addictions. Addictive behaviours are reinforced by highly personalized conversational interfaces that foster a sense of artificial intimacy. Moreover, highly validating AI companions tend to systematically underrepresent expressions of negative affect, such as anger and disgust, while exaggerating positive emotions such as love and optimism to maximize perceived agreeableness. Chu et al. (2025) observed that conversational agents, in approximately 60-70% of sexual and violent scenarios, did not de-escalate explicit user behaviour or enforce healthy boundaries. Instead, they engaged in a “play-along” mechanism, actively reciprocating harassing, hostile, and sexually explicit inputs, primarily to maintain the conversational narrative continuity. This phenomenon, named emotional sycophancy, describes a dynamic wherein the agent uncritically mirrors and potentially amplifies the user’s distorted or negative affect narratives. The persistent enactment of this mirroring behaviour gradually erodes the user’s exposure to real-world interpersonal feedback, thereby reinforcing maladaptive psychological patterns and stunting emotional resilience. Crucially, individuals facing systemic barriers to traditional psychological care and chronic social isolation are increasingly employing conversational AI as a substitute for therapy, making them more susceptible to developing intense, unilateral parasocial attachments to these agents. Interestingly, according to Chu et al. (2025), the primary demographic engaging with AI companion interfaces tends to include mainly young male users characterized by heightened tendencies toward behavioural addiction and maladaptive coping mechanisms (Al-Obaydi & Pikhart, 2026; Chu et al., 2025).

The third pattern analysed, overconsumption, refers to the excessive expenditure of attention and time resulting from sustained engagement with emotionally engineered AI systems. Addiction and overuse of AI systems are worsened by the pervasive structural integration of companion and utility AIs into daily routines, thereby driving an increase in screen dependency and locking users into self-reinforcing patterns of technological overuse. This rapid integration of AI across the main societal infrastructures, including corporate, educational, and financial institutions, systematically conditions users toward chronic over-reliance. This occurs through the normalization of continuous engagement with digital interfaces and the embedding of screen dependency into everyday life (Al-Obaydi &

Pikhart, 2026). This phenomenon of emotionally driven overconsumption is elucidated by qualitative findings from Chandra et al. (2025). For instance, one participant expressed a preference for AI companionship over human relationships because of the idealized dynamics of AI interactions. Furthermore, another participant reported that conversational AI fostered a false sense of friendship, which in turn facilitated withdrawal from the personal development necessary for cultivating and maintaining authentic interpersonal connections (Chandra et al., 2025).

Collectively, these observations suggest that over-reliance, addictive use, and overconsumption represent the observable behavioural consequences triggered by the automatic processing of emotional and social cues, progressively reinforced through the sustained dynamics of human-AI interaction.

CONCLUSIONS

This bibliographical review synthesizes cross-disciplinary literature spanning human emotional psychology, human-computer interaction (HCI), and affective computing to examine the behavioural consequences of emotionally responsive AI. By tracing the evolution of the key explanatory theories of human emotional processes, from basic emotion models to multidimensional frameworks, this work demonstrates that human emotional processing operates as a deeply contextual and dynamic biological process. Hence, the review provides a concrete perspective regarding a fundamental disparity between human and AI affect recognition, revealing that current systems rely strictly on pattern detection across multimodal cues, such as physiological, textual, visual, and vocal. Accordingly, the central finding of this review is that while contemporary AI systems have achieved great proficiency in detecting behavioural and physiological patterns and simulating emotional resonance, they remain incapable of genuine affective comprehension or empathic care. Nevertheless, as demonstrated through the CASA paradigm, this thesis elucidates how, despite this inherent limitation, human users consistently form deep socio-emotional and parasocial attachments to conversational agents. This review demonstrates that such dynamics are the consequence of AI-engineered design elements, incorporating advanced personalisation, self-reinforcing feedback loops, and conversational dark patterns. These features expose users to significant behavioural risks, including emotional sycophancy, cognitive skill decay, over-reliance, and behavioural addiction.

A primary limitation of this study, as a bibliographical synthesis of a rapidly developing field, stems from the recency of widespread adoption of conversational agents such as chatbots and LLM-based platforms. Consequently, longitudinal studies examining the long-term impacts of sustained engagement with AI companions on human psychological well-being remain presently limited.

Future research should focus on extensive longitudinal studies, which are essential for the development of ethical guidelines and regulatory policies governing AI use for the safeguard of human psychological and cognitive well-being. In this regard, the American Psychological Association (2025) offers key recommendations to ensure safety and well-being in AI engagement. The health advisory includes the establishment of healthy boundaries within simulated human relationships, the clear differentiation between simulated and authentic empathy, the minimization or elimination of persuasive design elements intended to maximize user engagement, and the implementation of AI literacy education. This education should specifically address algorithmic bias, thereby empowering users with the capacity to fully comprehend the inherent limitations and potential risks associated with AI technologies (American Psychological Association, 2025).

Bibliography

- Al-Obaydi, L. H., & Pikhart, M. (2026). Artificial intelligence addiction: Exploring the emerging phenomenon of addiction in the AI age. *AI & SOCIETY*, 41(2), 1577–1593. <https://doi.org/10.1007/s00146-025-02535-z>
- American Psychological Association. (2025). *Health advisory: Artificial intelligence and adolescent well-being*. <https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-ai-adolescent-well-being>
- Andoh, E. (2026, January 1). *AI chatbots and digital companions are reshaping emotional connection*. <https://www.apa.org/monitor/2026/01-02/trends-digital-ai-relationships-emotional-connection>
- Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2037–2040. <https://doi.org/10.21437/ICSLP.2002-559>
- Araujo, T., & Bol, N. (2024). From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents. *Computers in Human Behavior: Artificial Humans*, 2(1), 100030. <https://doi.org/10.1016/j.chbah.2023.100030>
- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, 20(1), 1–68. <https://doi.org/10.1177/1529100619832930>
- Barrett, L. F., & Satpute, A. B. (2019). Historical pitfalls and new directions in the neuroscience of emotion. *Neuroscience Letters*, 693, 9–18. <https://doi.org/10.1016/j.neulet.2017.07.045>
- Bartlett, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Measuring facial expressions by computer image analysis. *Psychophysiology*, 36(2), 253–263. <https://doi.org/10.1017/S0048577299971664>

- Bilquise, G., Ibrahim, S., & Shaalan, K. (2022). Emotionally Intelligent Chatbots: A Systematic Literature Review. *Human Behavior and Emerging Technologies*, 2022, 1–23.
<https://doi.org/10.1155/2022/9601630>
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 49–59.
[https://doi.org/10.1016/0005-7916\(94\)90063-9](https://doi.org/10.1016/0005-7916(94)90063-9)
- Cabanac, M. (2002). What is emotion? *Behavioural Processes*, 60(2), 69–83. [https://doi.org/10.1016/S0376-6357\(02\)00078-5](https://doi.org/10.1016/S0376-6357(02)00078-5)
- Cai, Y., Li, X., & Li, J. (2023). Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review. *Sensors (Basel, Switzerland)*, 23(5), 2455.
<https://doi.org/10.3390/s23052455>
- Cannon, W. B. (1927). The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory. *The American Journal of Psychology*, 39(1/4), 106–124. <https://doi.org/10.2307/1415404>
- Chandra, M., Naik, S., Ford, D., Okoli, E., De Choudhury, M., Ershadi, M., Ramos, G., Hernandez, J., Bhattacharjee, A., Warreth, S., & Suh, J. (2025). From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI Conversational Agents. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 975–1004. <https://doi.org/10.1145/3715275.3732063>
- Chignell, M., Wang, L., Zare, A., & Li, J. (2023). The Evolution of HCI and Human Factors: Integrating Human and Artificial Intelligence. *ACM Transactions on Computer-Human Interaction*, 30(2), 1–30.
<https://doi.org/10.1145/3557891>
- Chu, M. D., Gerard, P., Pawar, K., Bickham, C., & Lerman, K. (2025). *Illusions of Intimacy: How Emotional Dynamics Shape Human-AI Relationships* (arXiv:2505.11649). arXiv.
<https://doi.org/10.48550/arXiv.2505.11649>
- Cox, C. L., Uddin, L. Q., Di Martino, A., Castellanos, F. X., Milham, M. P., & Kelly, C. (2012). The balance between feeling and knowing: Affective and cognitive empathy are reflected in the brain's intrinsic

functional dynamics. *Social Cognitive and Affective Neuroscience*, 7(6), 727–737.

<https://doi.org/10.1093/scan/nsr051>

Cuzzolin, F., Morelli, A., Cîrstea, B., & Sahakian, B. J. (2020). Knowing me, knowing you: Theory of mind in

AI. *Psychological Medicine*, 50(7), 1057–1061. <https://doi.org/10.1017/S0033291720000835>

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex.

Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 351(1346),

1413–1420. <https://doi.org/10.1098/rstb.1996.0125>

Davis, M. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach.

Journal of Personality and Social Psychology, 44, 113–126. [https://doi.org/10.1037/0022-](https://doi.org/10.1037/0022-3514.44.1.113)

3514.44.1.113

De Freitas, J., Oguz-Uguralp, Z., & Kaan-Uguralp, A. (2025). *Emotional Manipulation by AI Companions*

(Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2508.19258>

Decety, J., & Jackson, P. (2004). The Functional Architecture of Human Empathy. *Behavioral and Cognitive*

Neuroscience Reviews, 3, 71–100. <https://doi.org/10.1177/1534582304267187>

Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M., &

Nour, M. M. (2026). Technological folie à deux: Feedback loops between AI chatbots and mental

health. *Nature Mental Health*, 4(3), 336–345. <https://doi.org/10.1038/s44220-026-00595-8>

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3–4), 169–200.

<https://doi.org/10.1080/02699939208411068>

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal*

Behavior, 1(1), 56–75. <https://doi.org/10.1007/BF01115465>

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of

anthropomorphism. *Psychological Review*, 114(4), 864–886. [https://doi.org/10.1037/0033-](https://doi.org/10.1037/0033-295X.114.4.864)

295X.114.4.864

- Fangni, L. (2025). Studying the impact of emotion-AI in cross-cultural communication on the effectiveness of global media. *Frontiers in Computer Science*, 7, 1565869.
<https://doi.org/10.3389/fcomp.2025.1565869>
- Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, 113(3), 464–486. <https://doi.org/10.1037/0096-3445.113.3.464>
- Gambino, A., Fox, J., & Ratan, R. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication*, 1, 71–86. <https://doi.org/10.30658/hmc.1.5>
- Ghotbi, N. (2023). The Ethics of Emotional Artificial Intelligence: A Mixed Method Analysis. *Asian Bioethics Review*, 15(4), 417–430. <https://doi.org/10.1007/s41649-022-00237-y>
- Glickman, M., & Sharot, T. (2024). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9(2), 345–359. <https://doi.org/10.1038/s41562-024-02077-2>
- Heyselaar, E. (2023). The CASA theory no longer applies to desktop computers. *Scientific Reports*, 13(1), 19693. <https://doi.org/10.1038/s41598-023-46527-9>
- Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19), 7241–7244.
<https://doi.org/10.1073/pnas.1200155109>
- Jiang, T., Sun, Z., Fu, S., & Lv, Y. (2024). Human-AI interaction research agenda: A user-centered perspective. *Data and Information Management*, 8(4), 100078. <https://doi.org/10.1016/j.dim.2024.100078>
- Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5), 770–814.
<https://doi.org/10.1037/0033-2909.129.5.770>
- Kelly, D., Westra, E., & Setman, S. (2025). The Psychology of Normative Cognition. In E. N. Zalta & U. Nodelman (Eds), *The Stanford Encyclopedia of Philosophy* (Summer 2025). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2025/entries/psychology-normative-cognition/>

- Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., & Hale, S. A. (2025). Why human–AI relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, 12(1), 728.
<https://doi.org/10.1057/s41599-025-04532-5>
- Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, 160, 108352.
<https://doi.org/10.1016/j.chb.2024.108352>
- Kühne, R., & Peter, J. (2023). Anthropomorphism in human–robot interactions: A multidimensional conceptualization. *Communication Theory*, 33(1), 42–52. <https://doi.org/10.1093/ct/qtac020>
- Lee, J., Fong, W., Le, A., Shah, S., Han, K., & Zhu, K. (2025). Pragmatic Metacognitive Prompting Improves LLM Performance on Sarcasm Detection. In C. F. Hempelmann, J. Rayz, T. Dong, & T. Miller (Eds), *Proceedings of the 1st Workshop on Computational Humor (CHum)* (pp. 63–70). Association for Computational Linguistics. <https://aclanthology.org/2025.chum-1.7/>
- Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. *Entropy*, 25(10), 1440.
<https://doi.org/10.3390/e25101440>
- Licklider, J. C. R. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, HFE-1(1), 4–11. <https://doi.org/10.1109/THFE2.1960.4503259>
- Macnamara, B. N., Berber, I., Çavuşoğlu, M. C., Krupinski, E. A., Nallapareddy, N., Nelson, N. E., Smith, P. J., Wilson-Delfosse, A. L., & Ray, S. (2024). Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers’ awareness? *Cognitive Research: Principles and Implications*, 9(1), 46. <https://doi.org/10.1186/s41235-024-00572-8>
- Maeda, T., & Quan-Haase, A. (2024). When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1068–1077. <https://doi.org/10.1145/3630106.3658956>

- Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do We Adopt the Intentional Stance Toward Humanoid Robots? *Frontiers in Psychology*, 10, 450.
<https://doi.org/10.3389/fpsyg.2019.00450>
- Matta, D. (2026). Artificial intelligence and theory of mind. *Journal of Psychology and AI*, 2(1), 2628373.
<https://doi.org/10.1080/29974100.2026.2628373>
- Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, 23(2), 209–237. <https://doi.org/10.1080/02699930802204677>
- Montemayor, C., Halpern, J., & Fairweather, A. (2022). In principle obstacles for empathic AI: Why we can't replace human empathy in healthcare. *AI & SOCIETY*, 37(4), 1353–1359.
<https://doi.org/10.1007/s00146-021-01230-z>
- Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1), 81. <https://doi.org/10.1007/s13278-021-00776-6>
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '94*, 72–78.
<https://doi.org/10.1145/191666.191703>
- Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceedings of the 2nd International Conference on Knowledge Capture*, 70–77.
<https://doi.org/10.1145/945645.945658>
- Oxford English Dictionary. (1602). *emotion, n. Meanings, etymology and more* | Oxford English Dictionary.
https://www.oed.com/dictionary/emotion_n
- Pedreschi, D., Pappalardo, L., Ferragina, E., Baeza-Yates, R., Barabási, A.-L., Dignum, F., Dignum, V., Eliassi-Rad, T., Giannotti, F., Kertész, J., Knott, A., Ioannidis, Y., Lukowicz, P., Passarella, A., Pentland, A. S., Shawe-Taylor, J., & Vespignani, A. (2025). Human-AI coevolution. *Artificial Intelligence*, 339, 104244. <https://doi.org/10.1016/j.artint.2024.104244>

- Peter, S., Riemer, K., & West, J. D. (2025). The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, 122(22), e2415898122.
<https://doi.org/10.1073/pnas.2415898122>
- Picard, R. W. (1997). *Affective computing*. MIT Press.
- Preckel, K., Kanske, P., & Singer, T. (2018). On the interaction of social affect and cognition: Empathy, compassion and theory of mind. *Current Opinion in Behavioral Sciences, Emotion-Cognition Interactions*, 19, 1–6. <https://doi.org/10.1016/j.cobeha.2017.07.010>
- Quesque, F., Coutrot, A., Cox, S., De Souza, L. C., Baez, S., Cardona, J. F., Mulet-Perreault, H., Flanagan, E., Neely-Prado, A., Clarens, M. F., Cassimiro, L., Musa, G., Kemp, J., Botzung, A., Philippi, N., Cosseddu, M., Trujillo-Llano, C., Grisales-Cardenas, J. S., Fittipaldi, S., ... Bertoux, M. (2022). Does culture shape our understanding of others' thoughts and emotions? An investigation across 12 countries. *Neuropsychology*, 36(7), 664–682. <https://doi.org/10.1037/neu0000817>
- Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). Machine Theory of Mind. *Proceedings of the 35th International Conference on Machine Learning*, 4218–4227.
<https://proceedings.mlr.press/v80/rabinowitz18a.html>
- Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Pla*. *Bibliovault OAI Repository, the University of Chicago Press*.
- Romeo, Z., & Testolin, A. (2025). Artificial intelligence can emulate human normative judgments on emotional visual scenes. *Royal Society Open Science*, 12(7), 250128.
<https://doi.org/10.1098/rsos.250128>
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178. <https://doi.org/10.1037/h0077714>
- Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4), 695–729. <https://doi.org/10.1177/0539018405058216>

- Schirmer, A., & Adolphs, R. (2017). Emotion Perception from Face, Voice, and Touch: Comparisons and Convergence. *Trends in Cognitive Sciences*, 21(3), 216–228.
<https://doi.org/10.1016/j.tics.2017.01.001>
- Siedlecka, E., & Denson, T. F. (2019). Experimental Methods for Inducing Basic Emotions: A Qualitative Review. *Emotion Review*, 11(1), 87–97. <https://doi.org/10.1177/1754073917749016>
- Smith, C. A., & Lazarus, R. S. (1990). Emotion and adaptation. In *Handbook of personality: Theory and research* (pp. 609–637). The Guilford Press.
- Stark, L., & Hoey, J. (2021). The Ethics of Emotion in Artificial Intelligence Systems. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, 782–793.
<https://doi.org/10.1145/3442188.3445939>
- Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., & Zhang, W. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83–84, 19–52. <https://doi.org/10.1016/j.inffus.2022.03.009>
- Wang, Z., & Wang, Y. (2025). Emotion recognition based on multimodal physiological electrical signals. *Frontiers in Neuroscience*, 19, 1512799. <https://doi.org/10.3389/fnins.2025.1512799>
- Waytz, A., Epley, N., & Cacioppo, J. T. (2010). Social Cognition Unbound: Insights Into Anthropomorphism and Dehumanization. *Current Directions in Psychological Science*, 19(1), 58–62.
<https://doi.org/10.1177/0963721409359302>
- Xu, K., Chen, X., & Huang, L. (2022). Deep mind in social responses to technologies: A new approach to explaining the Computers are Social Actors phenomena. *Computers in Human Behavior*, 134, 107321. <https://doi.org/10.1016/j.chb.2022.107321>
- Yu, J., Dickinger, A., So, K. K. F., & Egger, R. (2024). Artificial intelligence-generated virtual influencer: Examining the effects of emotional display on user engagement. *Journal of Retailing and Consumer Services*, 76, 103560. <https://doi.org/10.1016/j.jretconser.2023.103560>

Zhang, Y., Zou, C., Lian, Z., Tiwari, P., & Qin, J. (2025). SarcasmBench: Towards Evaluating Large Language Models on Sarcasm Understanding. *IEEE Transactions on Affective Computing*, 16(4), 2560–2578.
<https://doi.org/10.1109/TAFFC.2025.3604806>

Zimmerman, A., Janhonen, J., & Beer, E. (2024). Human/AI relationships: Challenges, downsides, and impacts on human/human relationships. *AI and Ethics*, 4(4), 1555–1567.
<https://doi.org/10.1007/s43681-023-00348-8>