TOWARD EMOTIONALLY INTELLIGENT AI SYSTEMS: FOUNDATIONS, CHALLENGES, AND THE ROAD AHEAD

HEIDRICH VICCI [a*]

Abstract:

Artificial Intelligence (AI) systems are increasingly used in settings in which their outputs alter decisions, communication, and access to services. In such settings, technical accuracy does not exhaust the conditions of a satisfactory interaction, because a user's emotional state may affect both interpretation and outcome. This paper examines the theoretical foundations of Emotional Intelligence (EI), the emergence of affective computing, and the present capabilities of emotionally aware AI across applied contexts. It brings together evidence from psychology, computer science, human-computer interaction, and ethics to ask a narrower question: which affective capacities can be implemented reliably, where do present systems fail, and which failures matter most for human-facing deployment? The review identifies technical and ethical constraints, organizes the most consequential research gaps, and proposes priorities for future development. Its central argument is that, in consequential human-machine interaction, systems should be able to identify affective cues, interpret them in context, and modulate their responses without mistaking statistical regularity for human understanding. Emotional competence, so understood, is not an ornamental addition to AI but one condition of socially useful and ethically defensible performance.

Keywords: Artificial Intelligence, Emotional Intelligence, Affective Computing, Empathy, Human-Computer Interaction, Ethics, Mental Health.

JEL Classification: O33, I12, J24.

[a] College Of Business, Florida International University, Miami, Florida, USA.
hvicci@fiu.edu.

**Corresponding Author*

Introduction

The question is no longer whether machines can display responses that appear sensitive to emotion, but whether such responses improve outcomes in the settings in which AI is actually used. Across healthcare, education, customer service, financial advising, and social assistance, the emotional conditions of an interaction can alter how technically correct information is understood and acted upon. A diagnostic system may identify a condition accurately yet communicate it in a manner that increases distress; a tutoring system may detect confusion yet fail to adapt its explanation; a companion robot may provide logistical support while remaining poorly attuned to the user's state. These are not merely matters of style. They affect trust, comprehension, adherence, and the quality of human-machine cooperation (Abdollahi et al., 2023; Boucher et al., 2021).

These deployment problems explain the sustained scholarly interest in affective computing: the study of systems that recognize, interpret, model, and respond to affective information. Picard's foundational account framed affect not as decorative interface behavior but as information relevant to perception, decision, learning, and interaction (Picard, 1997). The practical question, however, is not whether a machine can produce an emotional-looking response. It is whether the inference on which that response rests is valid, whether context has been represented adequately, and whether the response is appropriate to the user's interests. Human judgment, communication, and cooperation routinely occur under affective conditions; an AI system that treats such signals as irrelevant noise therefore operates with incomplete information. Yet the opposite error is equally serious: inferring more about a person's inner state than the available evidence can justify.

An analysis of affective AI requires drawing on theory of Emotional Intelligence (EI). EI refers to abilities to perceive emotion, use affective information for thinking, comprehend emotional processes, and manage emotions (Salovey & Mayer, 1990; Mayer et al., 2004). To apply this construct to artificial intelligence does not require postulating the existence of equivalent internal emotional states. Instead, it involves specifying functions, that is, detecting relevant signals, modeling their ambiguity, integrating the information into decision-making, and selecting responses based on contextual cues. Thus, recognizing patterns of data alone is insufficient, and a normative theory is required to analyze whether such

https://doi.org/10.70150/2026er02.

recognition constitutes evidence in favor of justified action. As such, the main contribution of this paper is a review of research on emotionally intelligent AI. In particular, we discuss theoretical background of EI and its historical connection to machine cognition; state-of-the-art technology in affective AI; applications to real-world domains such as mental health, education, and social robotics; problems associated with implementation, validation, and ethical concerns; identification of major open research questions based on the analysis above; and concluding remarks. Our objective is analytical synthesis, rather than an extensive survey. There is an important distinction to make upfront. Our premise here is narrower than a general notion that machines "become emotional." If an AI system relies on affective information, then validation and ethics must develop in lockstep due to the risks of misinformation or manipulation. Every stage discussed above is related to one question: Does affective information lead to better judgment and interactions when uncertainty, context, and abuse possibilities are taken into consideration?

A note on methodology is warranted before proceeding. This article is intended as a critical narrative review of the intersection of Emotional Intelligence (EI) and Artificial Intelligence (AI). Literature was identified through searches of academic databases including PubMed, Scopus, Web of Science, Google Scholar, and ACM Digital Library, using terms combining 'emotional intelligence,' 'affective computing,' 'emotion recognition,' 'empathy in AI,' 'human-computer interaction,' and 'AI ethics.' Because the article is a narrative rather than a systematic review, the search was used to construct an analytically relevant corpus rather than to claim exhaustive retrieval. Priority was given to peer-reviewed empirical studies, reviews, and theoretically foundational works published between 2018 and 2024, with earlier definitional contributions retained where necessary, including Salovey and Mayer (1990) and Picard (1997). Sources were retained when they addressed implementation, evaluation, or governance of affective capacities in AI; papers concerned with AI or EI only tangentially were not used to support core claims. The resulting synthesis draws across psychology, computer science, human-computer interaction, clinical medicine, and ethics and identifies both current knowledge and the most consequential unresolved questions. As a narrative review, the article also has limitations that should remain visible to the reader. The corpus is necessarily selective, the terminology of emotional AI varies across disciplines, and a fast-moving literature can change the balance of evidence after the search period. These limitations do not prevent analytical synthesis, but they counsel against treating absence from the reviewed corpus as proof that no relevant work exists. Exact search dates, full query strings, and screening counts should be added from the author's research log before submission if they are available; they should not be reconstructed retrospectively without documentary support.

1. Defining Emotional Intelligence

Formalizing the notion of Emotional Intelligence began in the early 1990s. In particular, Mayer, Salovey, and Caruso formalized the concept as an ability approach. Under this approach, EI represents a type of intelligence concerning emotion-related information instead of a broad personality trait (Salovey & Mayer, 1990; Mayer et al., 2004). The model consists of four branches: accurate perception of emotions; utilization of emotion to inform thoughts; comprehension of emotions and their transformations; and management of emotions in oneself and others. As applied to Artificial Intelligence (AI), this decomposition has consequences because it splits a broad term into concrete abilities. It permits these abilities to be specified and measured separately and avoids the fallacy of concluding that, because an AI can successfully classify emotions, it must be emotionally intelligent as a whole. Such reasoning would imply an equivalence between emotion detection and emotion-based reasoning. Instead, these two tasks differ substantially, despite being related to each other.

Another school of thought, represented chiefly by Goleman's mixed model, extends EI to cover self-awareness, self-regulation, motivation, empathy, and social skills (Goleman, 1995). While this extension of EI proved useful for organizations and education theory, any computational instantiation needs to consider these abilities independently of importing a personality trait. For our purpose, we need to identify which aspects of EI may be implementable: affect awareness, affect regulation, adaptive social response, and appropriate delegation of responsibility. In general terms, human literature connects EI with positive effects such as well-being and workplace productivity. But this does not imply applications of AI. Instead, affective AI should meet a criterion that matches implemented functions rather than associations that justify applications of EI to machines. This consideration highlights a distinction between AI applications and evaluations.

Affective artificial intelligence should instead be thought of as a functional research agenda, rather than proof that computers genuinely feel emotions. Specifically, affective computing studies how computational agents perceive, interpret, and respond to affective cues in perception and interaction (Picard, 1997; De Togni et al., 2021). For example, a system might detect sadness or anxiety and produce a reasonable reaction based on the detected emotional state. Such competence is testable through empirical benchmarks. By contrast, any claim of understanding the subjective content of these detections introduces an entirely separate problem, not implied by behavioral performance alone. While deployment considerations depend critically on reliability, calibration, and scope of application, no ordinary benchmark is capable of verifying genuine comprehension or feelings. For clarity, the terminology must clearly distinguish among emotion detection, affectively guided action, and any assertion of actual feeling on behalf of the computer. The distinction between these categories is essential, since they are separable both empirically and in principle.

2. Overview of Artificial Intelligence and Emotional Capacity

Our terminology distinguishes artificial intelligence into domain-restricted architectural solutions, generative models of high transfer ability, and speculative approaches to creating Artificial General Intelligence (AGI) (Cichocki & Kuleshov, 2021). Any taxonomy of narrow versus AGI AI should be used cautiously and only to avoid conflating different aspects of capability. For example, abilities such as emotion perception, affective inference, response regulation, social reasoning, and moral constraints can vary independently from generic intelligence. Just as Jeste et al. (2020) distinguish intelligence from a more comprehensive notion of wisdom, and Latif et al. (2023) explore the growing role of generality in educational contexts, we adopt the terminology here in order to evaluate affective abilities without reference to new ontological categories or automatic consequences thereof. The reason is a standard category mistake. Improved performance in natural language generation or logical inference does not necessarily imply the acquisition of specific affective competences. Our discussion focuses exclusively on affective capabilities as distinct from general computational performance.

A computational system's competence in detecting emotional cues differs fundamentally from its capacity for emotional intelligence. To detect an emotion is simply to perform statistical inference about facial expressions, tone of voice, written communication, physical behavior, or physiological responses. Emotional intelligence additionally involves an assessment of relevance and response appropriateness to the perceived emotional cue. For instance, while an inference engine may recognize emotions such as anger, anxiety, or confusion, its decision-making process provides little guidance on appropriate actions. The core question of interest lies elsewhere, namely, whether affective information is indeed relevant, how uncertainty affects judgments, and what guarantees prevent manipulation or exploitation. The leap from emotion detection to emotionally intelligent action amounts to a question of judgment.

Computational intelligence is employed widely in applications within manufacturing, health care, media, agriculture, security, finance, and transportation via predictive models, recommender systems, conversational AI, and robots (Zhou & Jiang, 2024). According to a 2017 PricewaterhouseCoopers report, AI could add up to US$15.7 trillion to worldwide GDP by 2030 (PwC, 2017). The cited figure is relevant to AI generally but does not serve as evidence of affective AI per se. Its relevance here is narrower: as AI moves into human-facing services, realized value depends not only on computational performance but also on whether users can understand, trust, and appropriately contest the system's behavior. Affective design can therefore influence adoption and service quality, but its economic value must be demonstrated rather than presumed.

3. Historical and Conceptual Foundations of Affective AI

Even earlier than contemporary advances in generative AI were made, affective computing research emerged from the recognition that successful machine interaction cannot rely solely on computational competence. Picard pioneered affective computing research by treating affect as informative signals and developing technologies to use them (Picard, 1997). The significance of Picard's contribution can also be attributed to restraint rather than boldness: she left unresolved whether machines should develop affective capabilities or instead respond adequately to human affect. With increasingly human-like interactions facilitated by advanced natural language processing and multimodal communication, distinguishing between these possibilities becomes crucial. Affect recognition can reasonably inform the design of machine responses, while attributing motives and feelings to the agent itself raises important ethical concerns. Recent literature indicates that partial implementation of affect sensitivity is technologically feasible; however, feasibility does not imply human equivalence.

The motivation behind this study stems from the recognition that emotions and cognition interact within human interaction. In this paper, we adopt a definition of emotional intelligence that includes perceiving, using, understanding, and managing affective information. The purpose of adopting this definition is to avoid the stronger hypothesis that a computer needs sentiments or consciousness in order to behave rationally in emotional contexts. Although debate about machine consciousness remains important, it should be distinguished from questions that can be tested directly: whether a system's affective inferences are valid, whether its responses are useful, and whether users are misled about the nature of the interaction (Weber-Guskar, 2021). There is empirical evidence showing that partial versions of affect-sensitive computer interaction are computationally feasible; however, computability is different from human equivalence. When an algorithm succeeds in simulating humans more convincingly, questions of disclosure, calibration, and limits on anthropomorphization become increasingly important. These concerns are particularly pressing when considering education.

For educational technology, affect matters because confusion, frustration, boredom, and confidence affect attention and perseverance. Recent literature on education-oriented AI recognizes that emotional and social interaction is important to adaptive tutoring applications, stressing at the same time the importance of deploying AI responsibly (Latif et al., 2023). The reasonable aim is complementarity rather than substitution: Computers offer scalability, continuity, and the capability of detecting regularities, whereas educators maintain contextual comprehension, accountability, and interpretation of a learner's situation in addition to whatever the computer sees. Affective AI is most defensible when it complements, rather than substitutes for, human capabilities.

3.1. Understanding Emotional Intelligence in Humans

Emotion and intelligence refer to distinct yet intersecting domains of human cognition. Research into Emotional Intelligence in psychology addresses precisely this intersection. According to the Ability Model by Mayer, Salovey, and Caruso, EI refers to the abilities regarding emotion-relevant information processing rather than representing a personality trait (Mayer et al., 2004). Empirical studies document correlations between measures of EI and socially meaningful outcomes; however, these correlations differ widely depending both on the specific EI measure used and its context of application. This aspect is crucial for AI: For a theory of interest to engineers, it must be clear precisely which quantity is supposed to be measured. Without this clarification, "emotional intelligence" could serve as an umbrella term covering distinct quantities that satisfy varying evidentiary standards and suffer distinct failure modes.

On the other hand, the Goleman model represents a hybrid model that incorporates cognitive abilities along with motivations, social skills, and behaviors (Goleman, 1995). The versatility of the Goleman model explains its success in practical applications such as education and organizational psychology; yet, the same versatility makes computational implementation challenging. Artificial intelligence cannot 'develop motivation' or become 'socially skilled' just because certain behaviors fall into these categories. Computational implementation requires specifying what functions the behavior performs, what evidence supports their inference, and when the performance qualifies as adequate. Age, social roles, cultural background, and environmental factors retain relevance, not due to the assumption of some invariant level of EI but since these factors may influence how affective signals are expressed and perceived.

In AI research, the problem consists not only of replicating aspects of human cognition but deciding which affective capacities may be implemented computationally without overstating their meaning. The founding definition by Picard (1997) describes affective computing as dealing with systems that perceive, interpret, or generate affective states. The value of this definition resides precisely in its allowance for a clear distinction between computing some capacity of affective interaction and simulating feelings or experience. Evaluation of emotion AI covers several levels, from sensing, representation, and inference to responses. The major challenge in implementing the computational analysis of affective interaction is posed by the inherently embodied, contextual, culturally mediated, and partly observable character of human affect. Algorithms rely on proxy variables that cannot perfectly represent the mental state. Thus, the essential requirement of the rigorous computational analysis of affective behavior consists of representing this uncertainty.

3.2. Importance of Emotional Intelligence in Human Interaction

EI is relevant to human social interactions since affect provides useful clues regarding intentionality, focus, uncertainty, and stakes (De Togni et al., 2021; Weber-Guskar, 2021). Analogously, the evaluation of the relevant capacity of affective computing includes asking whether affect-sensitive behaviors contribute to intelligibility, appropriate expectations of reliability, and context-relevant support without misleading perceptions about machine capability. From the practitioner perspective, this makes affective design both a matter of performance and ethics: a tool needs to become effective without being deceptive by virtue of its ability to appear emotionally intelligent.

Formally, we assume emotions to consist of several components. The seminal contribution of Ekman's basic-emotion theory consists of identifying a small set of emotion categories whose facial expressions follow certain stereotypical configurations (Ekman, 1992). One must stress that this categorization scheme does not imply a direct ability to recognize someone's inner state by looking at his face. As shown in a recent literature review by Barrett et al. (2019), individuals differ considerably in their facial movement patterns, and facial movements per se provide insufficient diagnostic information for determining someone's emotions. For artificial-intelligence-based affect recognition, this poses a substantial limitation since it rules out any possibility of establishing universal correspondences between facial expressions and emotions. In addition, expressions can involve combinations, masks, cultural interpretations, or other mental states rather than the assigned category. Systems for affect recognition and generation require considering contexts, estimating uncertainties, validating models on diverse samples, instead of using a simplistic mapping between a face's appearance and emotions. It follows that one cannot argue based on cross-cultural recognizability of emotion expression that an observer is able to deduce a universal emotional state merely from facial configurations. Including contextual variables into affective models is crucial since context is part of the object under consideration.

It turns out computationally challenging for machines to both recognize and generate natural affective behavior that would exhibit enough flexibility to match humans' behavior during social interactions (Pietikäinen & Silvén, 2022). The problem cannot simply be reduced to expanding the repertoire of possible emotional responses. First, affect itself may be indirect, strategically hidden, mixed, or conditioned on past communication exchanges between agents. Second, affective artificial intelligence is supposed to be assessed not in terms of facilitating authentic dialogues, but on the basis of improving the quality of the response relative to non-affective communication, while remaining trustworthy and transparent and maintaining users' control rights. Emotional intelligence entails detecting affective expressions, interpreting them correctly under uncertainty, selecting proper behavioral responses, and receiving feedback on these decisions. The problem is especially difficult when applying emotional intelligence to conversation management applications. Anger, fear, frustration, and distress affect what counts as adequate aid; however, affective responses have to involve distinguishing between registering cues, adapting messages' contents, and escalating an exchange due to exceeding one's capabilities.

4. From Affective Recognition to Emotionally Informed Agency

More powerful AI systems have not resolved the challenge of emotional intelligence; they have changed its technical form. Contemporary affective AI benefits from predictive models and multimodal signals, but additional capability does not automatically produce sound affective judgment. Experiments with large language models show that risk and prosocial cues can alter response patterns without implying that the model itself experiences emotion (Zhao et al., 2024). This distinction is instructive: sensitivity to affectively relevant input may improve adaptation, yet the interpretation of that input and the policy governing the response remain separate design problems. Emotionally informed agency therefore requires more than decoding expression; it requires a defensible relation between evidence, inference, and action.

Human EI theories were developed for embodied agents whose affective states are grounded in physiology, memory, social learning, and lived experience. Artificial systems do not share that basis, and this difference makes direct analogies hazardous. Bias can enter through sampling, annotation, measurement, and the objectives used to optimize behavior. The relevant question is therefore not merely whether a machine lacks a human body, but which functions of embodied experience are lost when affect is represented through observable proxies and learned correlations.

That limitation does not make affective AI impossible; it makes its claims conditional. In multimodal robots and conversational systems, deep learning, computer vision, speech, and language models can be combined to infer affective cues and adapt behavior. Such systems may be useful in companionship, education, or wellbeing support, but their outputs remain functional constructions rather than evidence of felt experience. Work on visual emotion analysis underscores both the expanding technical repertoire and the continuing ethical implications of affective inference (Wang et al., 2023).

Two broad development paths can be distinguished. One extends predefined emotion taxonomies and adds modules that recognize additional signals or select additional responses; the other learns affective representations with fewer predefined categories and greater dependence on data. The former is easier to audit but may freeze oversimplified categories; the latter can capture richer patterns but makes validation and normative control more difficult. Neither approach removes the prior question of purpose: which human interests should the system serve, what errors are tolerable, and who is accountable when an apparently empathetic response produces harm (Wang et al., 2023).

4.1. Integration of Emotional Intelligence in AI Systems

In research on empathetic AI, emotion recognition, affect representation, belief modeling, and response generation are often implemented as distinct components whose outputs are combined downstream. This modular separation is analytically useful because success at one stage does not guarantee success at the next. Accurate recognition is a necessary input in many applications, but an inference about a user's state becomes valuable only when the system can reason about its relevance, preserve uncertainty, and choose an action compatible with the user's interests. Empathetic behavior therefore cannot be reduced to a high classification score.

The integration of EI matters where an AI system must act on information about another person's state. In empathetic planning, for example, beliefs and affective information about a human partner can alter the selection of plans and assistance (Shvo & McIlraith, 2019). This creates a second-order problem: the system must decide not only what the user may be feeling but how that inference should constrain planning under deadlines, resource limits, and incomplete knowledge. In collaborative settings, the defensible goal is not to reproduce a human emotional persona; it is to make affective information relevant where it improves coordination and to disregard it where it would distort judgment.

Empathetic AI has attracted interest in education, healthcare, social robotics, and digital mental health. Chatbots and assistive systems can provide continuity, structured support, and rapid response, and qualitative studies of mental-health applications show that users can experience such tools as useful while also identifying limits in personalization and relational depth (Malik et al., 2022). The possibility that an affective interface can influence behavior is precisely why beneficial intent is insufficient as an ethical criterion. Steering should be transparent, proportionate to the user's own goals, open to refusal, and constrained against exploiting vulnerability. An emotionally persuasive system that obscures those conditions may increase engagement while reducing autonomy.

4.2. Advancements in Emotion Recognition and Understanding

With growing volumes of behavioral, textual, audiovisual, and physiological data, deep learning has become an important technology for affect recognition (Zhang et al., 2023; Kaklauskas et al., 2022). Ensemble methods and multimodal fusion can combine complementary channels, while regularization and careful validation are needed to limit overfitting. Yet speed and scale should not be confused with validity: a model can process signals more consistently than a human observer and still encode systematic measurement error or dataset bias. The relevant methodological question is whether performance generalizes across persons, cultures, devices, and contexts in which the distribution of signals differs from the training set.

Deep learning has produced substantial advances in emotion recognition and modeling over the past decade. Video and multimodal data are especially important because affect often unfolds over time, although frame-based models can remain competitive on some benchmarks. Convolutional, temporal-convolutional, recurrent, and transformer-based methods each offer tradeoffs in computation, interpretability, and sensitivity to dynamics. Temporal Convolutional Networks and Long Short-Term Memory models are among the architectures used for sequential affective signals, but no architecture is uniformly preferred across facial, bodily, vocal, and physiological modalities; relative performance depends on the task, the dataset, and the evaluation protocol (Kaklauskas et al., 2022; Zhang et al., 2023).

Many benchmark datasets are organized around prototypical facial categories such as anger, sadness, happiness, surprise, disgust, fear, and neutral states. That design has practical advantages for supervised learning, but it can reward performance on simplified labels that do not capture the ambiguity of everyday affective communication. Demographic imbalance, constrained recording conditions, label disagreement, and limited ecological validity can all restrict generalization. The appropriate response is not merely to enlarge datasets, but to document how labels were obtained, test performance across subgroups and settings, and report uncertainty when expressions do not fit a prototype (Barrett et al., 2019; Kaklauskas et al., 2022).

5. Empirical Landscape and Unresolved Research Problems

The contemporary landscape of AI and emotional intelligence research is active but fragmented, with persistent gaps between theoretical aspiration and empirical validation. Rather than a single mature field, affective AI encompasses partially overlapping traditions ranging from emotion recognition to social robotics, conversational agents, and human-computer interaction (HCI) (Weber-Guskar, 2021). NLP plays a prominent role in this field through the extraction of affective content and the conditioning of generation by inferred affective states. Several recent papers in healthcare apply combined NLP and deep learning approaches to detect sentiment and emotion-related content in patient-generated and clinical records (Nag et al., 2023). The promise is substantial, but the task is demanding: language can refer to affect indirectly, strategically, or ambiguously, and the confidence predicted by AI models should not exceed the actual informativeness of the input text.

Research on the COVID-19 period illustrates the uneven rather than absent state of the literature. Rezapour and Elmshaeuser (2022) used machine-learning and statistical models to examine factors associated with college students' emotional wellbeing during the pandemic and online learning. Such studies show that AI methods can organize complex behavioral data, but they do not by themselves establish that automated emotional inference should be used for diagnosis or intervention. Broader reviews of AI in mental health likewise emphasize potential benefits alongside privacy, bias, clinical-validation, and governance concerns (Olawade et al., 2024). The research need is therefore not simply more prediction, but evidence that prediction improves decisions and outcomes under realistic conditions.

5.1. Research Studies on Emotional Intelligence in AI

EI is relevant whenever AI needs to reason about the affective content of human actions and adjust its own behavior accordingly. For generative AI applications, affective context influences the acceptability of outputs, whereas personalization tasks involve adjusting timing, justification, and engagement according to perceived preferences. In these cases, distinguishing adaptive responses based on observable preferences from interpretations of the internal emotional states of users is particularly important. An AI system dealing with affect should treat the detected emotional content probabilistically, recognizing its provisional nature. In personalization tasks, this means adapting only to affective signals that are sufficiently reliable and relevant. Additionally, affective personalization should avoid assumptions based on stereotypes about certain populations' psychological features. For higher-stakes contexts such as medicine, finance, law, and organizational management, the requirements for affective systems are stricter: reasoning about any topic requires evidence capable of external examination, regardless of any emotional considerations. The affective component might facilitate communication and coordination, but affective inferences do not replace the domain knowledge necessary for effective decision-making.

The decomposition into components described above does not impose any assumption concerning the implementation of affective AI. While affective computing often implies simulating the internal processes involved, emotional AI applies equally to computational modeling of affect without making claims regarding the underlying mechanisms' nature. The notion of empathy is similarly reconsidered in our proposal: evaluating whether an AI system empathizes with humans is impossible within the standard methodology for assessing machine behavior. Instead, the goal is to determine if the system demonstrates context-appropriate empathic responses while acknowledging their artificiality and refusing to exploit users' anthropomorphism tendencies (Weber-Guskar, 2021; Zhou & Jiang, 2024). A Turing-style test of appearance alone would therefore be insufficient; evaluating affective systems requires assessing their performance in light of calibration, quality, subgroup fairness, and effects on future interactions. This framework maintains the essential ideas of the EI literature without equating machines with people and focusing instead on which abilities they have, how they interact, and at which points they diverge from human-like behavior.

5.2. Key Findings and Insights from Existing Literature

Ethical concerns become acute when an affective inference changes how a person is treated. A misclassification then ceases to be a mere prediction error and becomes an input to persuasion, monitoring, access, or care. Ethical evaluation must therefore ask who is affected, what evidence the system uses, how errors are distributed, and whether the person can understand and contest the inference (Ong, 2021; UNESCO, 2022). Context is decisive: the same emotionally expressive response may be reassuring in one setting and coercive in another. Cultural background, prior experience, disability, age, and institutional power can alter both expression and interpretation, so systems should not assume a universal mapping from signal to emotional meaning.

Current systems can produce behavior that users interpret as emotionally responsive, but evidence for robust emotional understanding across contexts remains limited. The distinction between appearing emotional and being emotionally intelligent is therefore not a rhetorical subtlety; it separates observable performance from an unverified claim about internal comprehension (De Togni et al., 2021; Weber-Guskar, 2021). A virtual assistant or social robot may detect affective cues, express sympathy, and regulate its wording effectively, yet still fail when cues are ambiguous, culturally unfamiliar, strategically presented, or unrelated to the user's actual needs. Progress consequently requires both better signal processing and clearer theories of what counts as successful affective conduct in a non-human agent.

From an engineering perspective, the relevant objective is not to confer a human personality on the system but to improve the quality of decisions made under affective uncertainty. A system should be able to recognize when an emotional cue is relevant, adjust communication when that adjustment serves the user's goals, and refrain from intervention when the evidence is weak or the stakes require human judgment. It should also be evaluated over time: a response that appears supportive in a single exchange may produce dependence, miscalibrated trust, or repeated error across sustained interaction. Emotional intelligence in AI is therefore best assessed as disciplined adaptation under constraints rather than as a general claim of human-like feeling. This is particularly important in high-stakes advisory settings. A system may improve communication by recognizing frustration or uncertainty, yet the underlying medical, financial, legal, or managerial recommendation must remain answerable to domain evidence rather than to the emotional state it has detected.

6. Limitations and Challenges

The development of emotionally intelligent AI confronts technical, theoretical, social, and ethical challenges that must be separated if they are to be addressed. One persistent difficulty is opacity: complex models can make it hard to explain why a particular affective inference or response was produced, especially when large-scale systems are optimized for broad deployment (Nasir et al., 2024). But opacity is not the only problem. Generic agents may also be calibrated toward the median user and fail at precisely the individual or culturally specific cases in which emotional context matters most. A credible account of emotional capability must therefore state what signals are used, how uncertainty is represented, where validation has occurred, and which decisions the system is not authorized to make. Such limits are part of competence, not admissions of failure. The demand for transparency is therefore not exhausted by an explanation of model internals. Users and institutions also need operational transparency: what the system can infer, what evidence supports the inference, how the information changes an action, and where human review remains mandatory.

Developers require diverse datasets of images, recordings, text, and physiological or behavioral signals if they are to avoid oversimplified representations of affective experience. This requires explicit assumptions about affect representation, uncertainty, state transition, and response policy rather than an undefined 'emotional calculus.' In categorical models, a new observation is mapped to patterns learned from previous examples; in dimensional or learned representations, similarity may be encoded more continuously. Either way, the system inherits the structure and omissions of its data (Assunção et al., 2023). Prototypical cases may be handled well while ambiguous, culturally specific, masked, or individually idiosyncratic expressions remain difficult. The design task is therefore to make those limits visible and operational, not to conceal them behind a single accuracy score.

The development of AI models with an emotional component presents unresolved technical and ethical questions (De Togni et al., 2021). There is no settled account of what machine EI should include, how affective reasoning should be represented, or how short-term adaptation should be reconciled with stable behavior over time. Personality frameworks such as the Five Factor Model or HEXACO may inform persistent behavioral parameters, but they cannot substitute for perception, uncertainty estimation, or regulation of affective response. Scalable deployment therefore requires mechanisms that distinguish stable system characteristics from transient user-state inferences and that prevent rapid adaptation from producing inconsistent or manipulative conduct (Ong, 2021). In practice, emotional stability and dynamic adaptation pull in different directions. A system that never adapts may be insensitive to context, whereas one that adapts too readily may become inconsistent across encounters. Architecture and evaluation must therefore specify which features may change, over what time horizon, and under whose control.

6.1. Ethical Considerations in Developing Emotional AI

Ethical risk arises when affective inference changes how a person is treated. Diverse training data and subgroup evaluation are necessary because sensing and classification errors can distribute unevenly across populations; the problem is especially serious when emotionally adaptive systems are used in settings of vulnerability or asymmetric power. Research on affective systems in games, for example, illustrates how emotion elicitation, sensing, and adaptive response can raise linked concerns about privacy, transparency, and ownership (Melhart et al., 2024). More general AI ethics frameworks add the requirements of accountability, human oversight, and documented limitations (Nasir et al., 2024). Users should also know when they are interacting with an AI system, particularly where children, people in crisis, or cognitively vulnerable users may otherwise attribute human understanding to a simulated response. The ethical standard should follow the consequence of the inference. A low-stakes conversational adaptation and a decision affecting employment, credit, education, or clinical care need not be governed identically, even if both begin from the same technical act of estimating an affective state.

Developers and deployers, rather than users, bear the primary responsibility for designing interactions that do not exploit psychological vulnerability. Affective systems can influence behavior through wording, timing, personalization, and the selective presentation of information; such influence becomes manipulation when it obscures purpose, bypasses meaningful choice, or uses inferred vulnerability against the person (Wang et al., 2023). Ong (2021) accordingly emphasizes provable beneficence and responsible stewardship as distinct duties for developers and operators. Transparency should not be treated as an enemy of effective interaction. The design problem is to disclose the system's artificial nature, data use, and relevant uncertainty in a manner compatible with the user's ability to understand and act on that information. The same principle applies to disclosure. Informing users that they are interacting with an artificial system is necessary, but meaningful transparency also requires clarity about whether emotions are being inferred, for what purpose, how long such data are retained, and whether the inference can be challenged or ignored.

Regulation is beginning to convert some of these concerns into enforceable limits. Under the European Union's Artificial Intelligence Act, the use of AI systems to infer emotions in workplaces and educational institutions is prohibited, subject to exceptions for medical or safety reasons; where emotion-recognition systems are lawfully used, other provisions can classify them as high-risk and Article 50(3) requires affected persons to be informed (Regulation (EU) 2024/1689). The scope should therefore not be described as a general prohibition on emotion recognition in law enforcement or every high-stakes setting. UNESCO's Recommendation takes a different, non-binding approach: it calls for awareness of anthropomorphization and technologies that recognize or mimic human emotions, especially where children are involved, and for research on the effects of long-term interaction (UNESCO, 2022). Together these instruments show that affective AI is not merely a technical frontier. It creates institutional power over how emotions are inferred, represented, and acted upon, and that power requires limits proportionate to the vulnerability and consequences of the setting. The regulatory lesson is correspondingly specific: institutions should not treat an affective score as neutral observational data merely because it is produced automatically. The permissibility of its use depends on purpose, setting, legal basis, safeguards, and the consequences attached to the inference.

6.2. Accuracy and Reliability of Emotion Recognition Algorithms

Although real-time emotion-recognition algorithms can detect facial, vocal, or other affective patterns under controlled conditions, numerous factors complicate their reliability in real-world deployment (Pietikäinen & Silvén, 2022). Recognition performance can degrade as expressions depart from benchmark prototypes, are blended or masked, or occur under cultural and situational conditions not represented in training. Barrett et al. (2019) further caution that facial movements do not provide a context-free readout of a person's emotional state. Aggregate accuracy can therefore conceal the conditions under which a model fails. A rigorous evaluation should report subgroup performance, calibration, uncertainty, and sensitivity to context, rather than assuming that a visible expression maps directly onto a single emotion. Reliability should thus be evaluated beyond aggregate accuracy. Calibration, false-positive and false-negative patterns, subgroup performance, sensitivity to context, and robustness under distribution shift are all relevant when an emotional label can alter the treatment or advice a person receives.

AI cannot be said, on current evidence, to reproduce the richness of human emotional experience, even when it recognizes affective patterns accurately (Weber-Guskar, 2021). Machine-learning systems can classify features in faces, voices, physiological signals, and text, but their outputs remain inferences from measured data rather than direct access to a person's inner state (Kaklauskas et al., 2022). Affective interfaces can nonetheless influence users through feedback, wording, and adaptation. That influence should be studied in measurable terms: changes in behavior, trust, dependence, mood, or decision quality. The ethical priority is to preserve wellbeing and autonomy rather than maximize engagement, and to examine long-term effects rather than infer benign impact from a convincing single exchange (UNESCO, 2022). The reference to emotional contagion is relevant chiefly as a warning about interaction effects: an interface need not possess an emotion in order to shape the emotional conditions of a conversation. What matters for governance is the measurable influence on the user and the incentives under which that influence is exercised.

7. Gap Analysis

The literature is uneven rather than simply sparse. Recognition accuracy and benchmark performance are comparatively well studied, whereas the translation of affective measures into reliable design rules for consumer-facing systems remains less mature. The central gap concerns the validity of connections between the different layers of processing involved—perception, reasoning, response, and ultimately human-level outcomes—rather than the mere absence of work on emotional-intelligence capabilities. Abilities to detect affective cues, reason under uncertainty, infer relevance to a given scenario, act appropriately, and escalate decisions to humans receive inconsistent definitions throughout the literature, resulting in substantial empirical investigation activity, but relatively few guidelines for combining these capacities effectively. In particular, a practitioner's challenge lies in ensuring a successful transition from detecting an affective cue reliably to acting successfully based on it and producing favorable long-term consequences. Showing good classification performance on a benchmark dataset does not imply that deploying the system to analyze consumer data leads to improved relationship management, treatment efficacy, learning experience quality, or other intended benefits.

Consequently, conducting a gap analysis reveals that the field lacks precisely a unified framework covering detection, interpretation under uncertainty, decision-making based on affective information, long-term impact, and governance. Different frameworks address aspects of this pipeline separately. Their respective methods, metrics, and application domains often differ greatly, which is relevant due to potential misunderstandings. Namely, it could happen that practitioners perceive accurate classification as sufficient proof of an emotionally intelligent system despite the underlying definition differences. A proper validation framework requires explicitly stating what capability is to be demonstrated, by which metrics and according to which standards, what kinds of errors are considered critical, and under what conditions escalation to humans is mandated. A useful framework would therefore have to connect measurement with action: it should specify which affective variables are observed, how uncertain estimates are combined with other evidence, which response options are admissible, how outcomes are monitored, and when the system must defer to a person.

7.1. Identifying Current Gaps in Emotional AI Research

In contrast to this literature, our perspective considers only selected aspects of an extensively growing body of literature, whose development is necessarily fragmented due to rapid advances and diverging epistemic priorities. In any event, several fundamental gaps emerge that require discussion: Firstly, standardized methodology does not exist even within individual modalities. Secondly, while detection rates are prioritized in empirical research, issues concerning response strategies and their ethical implications are comparatively understudied. Lastly, longitudinal data are rare. Without a clear understanding of the impact of sustained interactions, any conclusions drawn based on controlled experimental paradigms remain fundamentally limited. As illustrated by research on empathic planning and UNESCO's Recommendation on the Ethics of Artificial Intelligence, which calls for research on the effects of long-term human-AI interaction, there are substantial gaps with respect to long-term evidence concerning the effects of repeated interactions (Shvo & McIlraith, 2019; UNESCO, 2022). It is crucial to emphasize that these are fundamentally different types of gaps. While some concern issues of quantification and estimation, others relate to external validity and governance considerations. They must therefore be addressed separately; treating them as equivalent would obscure the unique evidence requirements underlying each.

Several important priority gaps deserve particular attention: First, lack of consistent definitions for the concept of AI-based EI. Second, poor diversity and ecological validity of training data. Third, insufficient validation across cultures and population subgroups. Fourth, weak consideration of epistemic uncertainty in modeling. And lastly, absence of longitudinal evaluations of sustained interaction. Empirical studies of student wellbeing during online learning demonstrate the need for predicting specific contextual variables rather than abstracted emotion labels (Rezapour & Elmshaeuser, 2022). Future research projects should therefore pair engineering evaluations with user-centered assessments of welfare outcomes, error tolerance and recovery, calibrated trust relationships, and contestability of decision processes. Addressing the gap requires interdisciplinary collaboration between CS, Psychology, Medicine, Ethics, and Social Science; the combination of technical accuracy with societal legitimacy demands separate lines of evidence.

7.2. Areas for Future Exploration and Development

Progress requires at least an architecture, namely, a specification that distinguishes between affect detection, uncertainty representation, inference regarding relevance, response determination, and governance constraints, and that allows evidence in one stage to modify the others. There are four research streams: improving the accuracy and calibration of each component separately; validating performance across cultures and use cases in situ; assessing user outcomes longitudinally; and testing disclosure requirements, escalation procedures, auditability, and resilience against manipulation in governance studies. These streams should be linked rather than optimized independently, because a more accurate detector can still produce a worse system if its inference is used at the wrong time or for the wrong purpose (Ong, 2021; Thakkar et al., 2024).

Greater expressive sophistication should be matched by commensurate evidence. An affective system may exaggerate understanding and thereby increase misplaced trust just when precaution is needed. Iteration involves stating the affective abilities, testing the system in realistic conditions, analyzing ethical and distributive implications, involving end-users to evaluate failure modes, and refining the system prior to further deployment. The goal is neither to reproduce all human emotions accurately nor to simulate understanding. Rather, it is to provide reliable help under uncertainty, along with adequate transparency enabling both the users and regulators to identify adaptive competence versus persuasive simulation. This benchmark is stringent, but it is precisely what is required for affective AI to succeed widely without undermining genuine social fluency through false claims of understanding.

8. Conclusion

The ability to identify and name emotion can assist empathetic interaction, but it is not identical with empathy. A plausible label may still misunderstand the circumstances that gave rise to an expression, and an apparently caring response can be used to influence as well as to assist (Weber-Guskar, 2021). The case for emotional AI should therefore not rest on the assumption that greater affective capability is inherently beneficial. The narrower and stronger claim is that, where affective information is genuinely relevant to a person's welfare or to the quality of cooperation, AI may use that information under constraints that protect autonomy, disclose uncertainty, and limit manipulation. Human flourishing is not secured by emotional realism alone; it depends on the purposes for which the capability is used and on the institutions that govern its use. For this reason, the value of emotional labeling depends on the discipline with which it is used. A tentative inference that prompts clarification may assist understanding; the same inference, treated as a fact about the person and used to steer behavior without consent, can become a mechanism of manipulation.

What, then, would justify describing an AI system as emotionally competent? The evaluation should begin not from whether the machine appears to feel, but from the people affected by its behavior. The system's inferences should be valid for the populations and contexts in which it is deployed; its uncertainty should be visible; its responses should improve rather than merely appear to improve welfare; and users should be able to refuse, contest, or escalate consequential interactions (Thakkar et al., 2024). Questions about artificial consciousness can remain open without weakening these practical standards. Indeed, conceptual work on emotionalized AI warns against confusing simulated reciprocity with human mutuality (Weber-Guskar, 2021). The relevant standard is therefore not whether a machine appears to feel. It is whether its use of affective information improves judgment and welfare without concealing uncertainty, exploiting vulnerability, or claiming a comprehension it cannot demonstrate. This criterion supports rather than weakens the paper's central position. Emotional competence matters precisely because human-facing systems operate inside social practices in which errors of interpretation, misplaced trust, and asymmetric influence have consequences that cannot be repaired by technical accuracy alone.

Conflicts of interest
The author(s) states that there is no conflict of interests.

References

Abdollahi, H., Mahoor, M. H., Zandie, R., Siewierski, J., & Qualls, S. H. (2023). Artificial emotional intelligence in socially assistive robots for older adults: A pilot study. IEEE Transactions on Affective Computing, 14(3), 2020–2032. https://doi.org/10.1109/TAFFC.2022.3143803

Assunção, G., Castelo-Branco, M., & Menezes, P. (2023). Self-mediated exploration in artificial intelligence inspired by cognitive psychology. Frontiers in Neurorobotics, 17, 1127869. https://doi.org/10.3389/fnbot.2023.1127869

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

Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., Parks, A. C., & Zilca, R. (2021). Artificially intelligent chatbots in digital mental health interventions: A review. Expert Review of Medical Devices, 18(sup1), 37–49. https://doi.org/10.1080/17434440.2021.2013200

Cichocki, A., & Kuleshov, A. P. (2021). Future trends for human-AI collaboration: A comprehensive taxonomy of AI/AGI using multiple intelligences and learning styles. Computational Intelligence and Neuroscience, 2021, Article 8893795. https://doi.org/10.1155/2021/8893795

De Togni, G., Erikainen, S., Chan, S., & Cunningham-Burley, S. (2021). What makes AI 'intelligent' and 'caring'? Exploring affect and relationality across three sites of intelligence and care. Social Science & Medicine, 277, 113874. https://doi.org/10.1016/j.socscimed.2021.113874

Ekman, P. (1992). An argument for basic emotions. Cognition & Emotion, 6(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Goleman, D. (1995). Emotional intelligence. Bantam Books.

Jeste, D. V., Graham, S. A., Nguyen, T. T., Depp, C. A., Lee, E. E., & Kim, H.-C. (2020). Beyond artificial intelligence: Exploring artificial wisdom. International Psychogeriatrics, 32(8), 993–1001. https://doi.org/10.1017/S1041610220000927

Kaklauskas, A., Abraham, A., Ubarte, I., Kliukas, R., Luksaite, V., Binkyte-Veliene, A., Vetloviene, I., & Kaklauskiene, L. (2022). A review of AI cloud and edge sensors, methods, and applications for the recognition of emotional, affective and physiological states. Sensors, 22(20), 7824. https://doi.org/10.3390/s22207824

Latif, E., Mai, G., Nyaaba, M., Wu, X., Liu, N., Lu, G., Li, S., Liu, T., & Zhai, X. (2023). AGI: Artificial general intelligence for education. arXiv. https://doi.org/10.48550/arXiv.2304.12479

Malik, T., Ambrose, A. J., & Sinha, C. (2022). Evaluating user feedback for an artificial intelligence-enabled, cognitive behavioral therapy-based mental health app (Wysa): Qualitative thematic analysis. JMIR Human Factors, 9(2), e35668. https://doi.org/10.2196/35668

Mayer, J. D., Salovey, P., & Caruso, D. R. (2004). Emotional intelligence: Theory, findings, and implications. Psychological Inquiry, 15(3), 197–215. https://doi.org/10.1207/S15327965PLI1503_02

Melhart, D., Togelius, J., Mikkelsen, B., Holmgård, C., & Yannakakis, G. N. (2024). The ethics of AI in games. IEEE Transactions on Affective Computing, 15(1), 79–92. https://doi.org/10.1109/TAFFC.2023.3276425

Nag, P. K., Bhagat, A., Vishnu Priya, R., & Khare, D. K. (2023). Emotional intelligence through artificial intelligence: NLP and deep learning in the analysis of healthcare texts. In 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI) (pp. 1–7). IEEE. https://doi.org/10.1109/ICAIIHI57871.2023.10489117

Nasir, S., Khan, R. A., & Bai, S. (2024). Ethical framework for harnessing the power of AI in healthcare and beyond. IEEE Access, 12, 31014–31035. https://doi.org/10.1109/ACCESS.2024.3369912

Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with artificial intelligence: Current trends and future prospects. Journal of Medicine, Surgery, and Public Health, 3, 100099. https://doi.org/10.1016/j.glmedi.2024.100099

Ong, D. C. (2021). An ethical framework for guiding the development of affectively-aware artificial intelligence. In 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 1–8). IEEE. https://doi.org/10.1109/ACII52823.2021.9597441

Picard, R. W. (1997). Affective computing. MIT Press. https://doi.org/10.7551/mitpress/1140.001.0001

Pietikäinen, M., & Silvén, O. (2022). Challenges of artificial intelligence—From machine learning and computer vision to emotional intelligence. arXiv. https://doi.org/10.48550/arXiv.2201.01466

PwC. (2017). Sizing the prize: What's the real value of AI for your business and how can you capitalise? PwC.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). Official Journal of the European Union, L, 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Rezapour, M., & Elmshaeuser, S. K. (2022). Artificial intelligence-based analytics for impacts of COVID-19 and online learning on college students' mental health. PLOS ONE, 17(11), e0276767. https://doi.org/10.1371/journal.pone.0276767

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. Imagination, Cognition and Personality, 9(3), 185–211. https://doi.org/10.2190/DUGG-P24E-52WK-6CDG

Shvo, M., & McIlraith, S. A. (2019). Towards empathetic planning. arXiv. https://doi.org/10.48550/arXiv.1906.06436

Thakkar, A., Gupta, A., & De Sousa, A. (2024). Artificial intelligence in positive mental health: A narrative review. Frontiers in Digital Health, 6, 1280235. https://doi.org/10.3389/fdgth.2024.1280235

UNESCO. (2022). Recommendation on the ethics of artificial intelligence. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Wang, J. Z., Zhao, S., Wu, C., Adams, R. B., Jr., Newman, M. G., Shafir, T., & Tsachor, R. (2023). Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion. Proceedings of the IEEE, 111(10), 1236–1286. https://doi.org/10.1109/JPROC.2023.3273517

Weber-Guskar, E. (2021). How to feel about emotionalized artificial intelligence? When robot pets, holograms, and chatbots become affective partners. Ethics and Information Technology, 23(4), 601–610. https://doi.org/10.1007/s10676-021-09598-8

Zhang, Z., Peng, L., Pang, T., Han, J., Zhao, H., & Schuller, B. W. (2023). Refashioning emotion recognition modelling: The advent of generalised large models. arXiv. https://doi.org/10.48550/arXiv.2308.11578

Zhao, Y., Huang, Z., Seligman, M., & Peng, K. (2024). Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots. Scientific Reports, 14, 7095. https://doi.org/10.1038/s41598-024-55949-y

Zhou, Y., & Jiang, R. (2024). Advancing explainable AI toward human-like intelligence: Forging the path to artificial brain. arXiv. https://doi.org/10.48550/arXiv.2402.06673