Beyond the Black Box: Ethical and Theoretical Grounding in Affective Computing

Wajdi Zaghouani
Northwestern University in Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Affective computing, the development of systems that recognize, interpret, and simulate human emotions, has advanced rapidly through deep learning and multimodal fusion techniques. Yet this technological progress has significantly outpaced foundational understanding of what emotions are, how they function socially, and what it means for machines to simulate them. This position paper argues that current affective systems are critically flawed because they rely on correlational patterns rather than established psychological theories of affect, are trained on biased data that systematically fails to generalize across demographic groups, and operate within inadequate ethical and regulatory frameworks that cannot protect emotional privacy or prevent harm. Drawing on empirical work in affective science, computational fairness research, and philosophical accounts of emotional expression, we argue for a theory-first approach that integrates psychological models, mandates rigorous fairness auditing across intersectional demographics, treats affective data as a protected category requiring heightened safeguards, and recognizes fundamental limits to emotion recognition systems. Without such grounding, affective computing risks systematically encoding bias, enabling emotional manipulation, and eroding authentic human connections that define meaningful social experience.

Keywords: affective computing, emotion recognition, ethical AI, fairness in AI, emotional privacy, psychological theory

1. Introduction: The Emotional Black Box

Emotion recognition systems are being rapidly deployed across high-stakes contexts including healthcare, education, employment screening, law enforcement, and consumer marketing. The underlying technology has become increasingly sophisticated. Systems trained on large datasets can detect emotional states from facial expressions with accuracy rates exceeding 80% (Kaur et al., 2025), from speech patterns with specialized neural architectures (Mustaqeem and Kwon, 2020), and from physiological signals collected through wearables (Zhang et al., 2020). Machine learning models are being used to predict depression from social media (De Choudhury et al., 2013), assess patient engagement in clinical settings (Tang et al., 2024), monitor student affect in classrooms (Tang et al., 2024), and estimate consumer emotional responses to marketing materials (Bacic and Gilstrap, 2024).

The promise is compelling: emotion recognition could enable earlier mental health intervention, improve educational responsiveness, enhance human-computer interaction, and provide objective measures of subjective states. Yet beneath this promise lies a troubling reality that spans four interconnected domains: theoretical, fairness-related, privacy-related, and ethical.

First, the theory problem: emotion recognition systems treat emotions as data to be extracted rather than as complex psychological phenomena embedded in individual goals, cultural display norms, and social relationships. They rely on correlational patterns between facial configurations, acoustic features, or physiological signals and categorical emotion labels, without incorporating established theories of how emotions actually work (Barker et al., 2025).

Second, the fairness crisis: systems trained predominantly on Western populations systematically underperform on non-Western emotional expressions. When emotion recognition systems encounter populations not well-represented in training data, accuracy drops sharply (Fan et al., 2023), with documented performance gaps of 27.2 percentage points or more across racial groups (Fan et al., 2023). Performance disparities extend across age (Kim et al., 2021), gender (Chien et al., 2024), and language (Nakai et al., 2023), raising fundamental questions about whether systems can be "fair" if they are trained on structurally biased datasets.

Third, the privacy paradox: emotion recognition produces data of unprecedented intimacy. Unlike traditional personal data, emotional states cannot easily be concealed or controlled. Facial expressions, voice patterns, and physiological responses can reveal information involuntarily, and when continuous surveillance systems make such revelation systematic, they enable inferences about health, political views, sexual orientation, and other deeply personal characteristics (Fabiano, 2025). Yet existing privacy regulations do not adequately protect affective data.

Fourth, the simulation problem: systems mar-

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 235–243
16 May 2026.

keted as emotionally responsive are, at their core, performing sophisticated mimicry. They have no genuine understanding of human suffering, no capacity for empathy, no authentic concern for user wellbeing. When deployed with vulnerable populations, this simulation can cause deep harm by providing false reassurance of connection and care (Barker et al., 2025).

This position paper articulates these four problems in detail, situates them within the broader debate on responsible affective computing, and proposes specific paths forward. While earlier work, notably Barker et al. (2025), has catalogued ethical considerations in emotion recognition research, and fairness-oriented surveys have documented demographic performance gaps, our contribution lies in *integrating* these four dimensions, theory, fairness, privacy, and simulation, into a single analytical framework and grounding each in specific design principles and policy recommendations that go beyond what any single prior study has articulated. In particular, we draw on recent developments in affective psychology, including constructionist, appraisal, sociodynamic, and evolutionary theories (Barrett, 2014; Moors, 2014; Mesquita and Boiger, 2014; Tracy, 2014), to argue that the theoretical deficit is not merely a scholarly concern but the root cause of failures in the other three domains.

2. The Theory Problem: Toward Psychologically Grounded Affective Computing

2.1. Correlational Patterns Are Not Theoretical Understanding

Current affective computing systems achieve impressive accuracy by identifying patterns in training data that correlate with emotion labels. Yet correlation is not understanding. A system that predicts "sadness" with 85% accuracy because it has learned that downward-turned mouth corners and decreased eye activity correlate with sadness labels does not understand what sadness is, what causes it, how it functions in human life, or what role it plays in decision-making, memory formation, or social relationships.

This distinction matters because emotion labels in training datasets derive from human annotators making interpretive judgments. All such judgments are structured by what researchers call "Qualities of Indeterminacy": subjectivity (meaning varies with who judges), uncertainty (annotators lack confidence), ambiguity (multiple valid interpretations), and vagueness (categories operate at different specificity levels) (Barker et al., 2025). Our own work on hate speech annotation has demonstrated that annotators experience significant emotional and cognitive toll during labeling tasks, and that the subjective nature of annotation decisions can introduce systematic biases into datasets (Al Emadi and Zaghouani, 2024). When a training dataset contains video clips labeled "angry" by human judges, those labels already embed particular cultural assumptions about what anger looks like, how it should be expressed, and in what contexts it is appropriate.

2.2. Beyond Basic Emotion Theory: Insights from Affective Psychology

Psychological research has largely moved beyond theories assuming emotions are universal, discrete categories. Contemporary affective science offers several more nuanced frameworks, as comprehensively reviewed in a special issue of *Emotion Review* (Volume 6, Issue 4, 2014) that brought together leading theorists from constructionist, appraisal, sociodynamic, and evolutionary traditions. We summarize these below and discuss their implications for affective computing.

Appraisal Theory. As Moors (2014) outlines, appraisal theories hold that emotions arise from how individuals evaluate events relative to their goals, values, and coping capacity. Anger results from appraising an event as goal-blocking, unjust, and controllable; fear results from appraising threat and uncertainty about effective coping. No current machine learning system incorporates such appraisal structures, making it impossible for these systems to distinguish between superficially similar expressions that actually reflect different emotional meanings.

The Conceptual Act Theory. Barrett (2014) proposes that emotions are not pre-wired responses but are constructed moment-by-moment through the interplay of bodily sensations, prior experiences, cultural knowledge, and situational context. On this view, an emotion category like "anger" is a population of heterogeneous instances tailored to specific environments, not a single recognizable pattern. This framework fundamentally challenges the assumption that recognizing emotion is a matter of pattern matching, suggesting instead that emotion perception is an active interpretive process shaped by cultural knowledge and previous experience.

The Sociodynamic Model. Mesquita and Boiger (2014) argue that emotions are dynamic systems emerging from social interactions and relationships. In this model, sociocultural environments do not merely provide content for cognitive representations; they actively shape the emotional episode through affordances, constraints, and reward structures. This implies that decontextualized emotion recognition, the standard approach in affective computing, strips away precisely the social context that

gives emotional expressions their meaning.

Evolutionary Approaches. Tracy (2014) argues that emotions are shaped by natural selection to facilitate adaptive responses to threats and opportunities. While this view supports the existence of some cross-cultural commonalities in emotional expression, it also predicts significant variability, since the same emotion serves different functions in different environments. Tracy emphasizes that even within an evolutionary framework, emotional responding is expected to vary substantially across individuals and contexts.

Dimensional Models. These propose that emotions vary along dimensions like valence (positive/negative), arousal (high/low), and dominance (feeling in control/helpless) rather than falling into discrete categories (Zhang et al., 2020). Cross-cultural research reveals that the number of dimensions necessary to capture emotional variation differs across cultures. East Asian populations perceive emotional expressions as varying along fewer dimensions than Western populations, with happy, sad, and angry expressions constituting one larger category in some East Asian contexts (Chen et al., 2024; Saito et al., 2023).

Cultural Display Rules and Emotion Socialization. The relationship between internal emotional states and external expressions is heavily culturally shaped. Collectivist cultures tend to emphasize emotional restraint to preserve group harmony, leading to more muted facial expressions for the same internal states compared to individualist cultures (Pruessner and Altan-Atalay, 2025; Yuk et al., 2025). Some cultures cultivate direct emotional expression while others favor indirect communication of emotional content through contextual cues (Li et al., 2025). The same facial configuration can signal different emotional states across cultures, and the same emotional state can be expressed through radically different facial configurations (Chen et al., 2024; Christensen et al., 2025).

These theories are not merely academic abstractions. Each has direct implications for how affective computing systems should be designed, what their outputs should represent, and what limitations must be acknowledged. The failure of current systems to engage with any of these frameworks represents a fundamental gap, not just in scholarly rigor, but in engineering practice.

2.3. Design Principles for Theory-Grounded Systems

We propose that affective computing must adopt what we call a "theory-first" architecture.

First, systems should represent emotions along multiple dimensions rather than forcing single-label categorical classification. This requires moving away from output layers producing "angry" / "sad" / "happy" / "neutral" classifications toward continuous representations of valence, arousal, and other dimensions (Zhang et al., 2020). We acknowledge that moving to continuous representations introduces new annotation challenges, since continuous ratings may amplify subjective variability among annotators. Addressing this requires careful calibration procedures and modeling of annotator disagreement as informative signal rather than noise.

Second, systems must incorporate contextual information: not just what someone's face looks like, but what situation they are in, what goals they might have, what cultural norms apply (Mesquita and Boiger, 2014). This pushes toward multimodal systems that integrate scene understanding, discourse analysis, and relationship modeling rather than treating facial expressions as decontextualized signals.

Third, annotation practices must acknowledge indeterminacy rather than forcing false consensus (Barker et al., 2025). Rather than single labels per instance, best practice involves collecting multiple interpretations per moment and modeling the distribution of possible affective meanings. Some datasets are beginning to adopt such approaches; for example, work on modeling annotator disagreement in NLU tasks and the CrowdTruth framework have demonstrated the value of preserving label distributions. Most emotion recognition datasets, however, still collapse annotations into single gold labels.

Fourth, systems must be transparent about fundamental limits. Perfect emotion recognition may be impossible and may be undesirable. Emotions partly function through what might be called productive opacity: because we do not have complete access to others' internal states, we must engage in genuine interpretation, which enables trust, vulnerability, and authentic connection (Barker et al., 2025).

3. The Fairness Crisis: Demographic Disparities, Intersectionality, and Structural Bias

3.1. The WEIRD Dataset Problem and Demographic Bias

Affective computing systems are trained predominantly on data from Western, Educated, Industrialized, Rich, Democratic (WEIRD) populations. A comprehensive review of eight major facial expression datasets (RAVDESS, FER2013, AffectNet, CK+, JAFFE, SFEW, EmotioNet, RAF-DB) documents substantial demographic imbalances (Kaur et al., 2025). These datasets underrepresent diverse racial, ethnic, gender, age, and health

groups. When systems trained on demographically imbalanced data encounter populations underrepresented in training, systematic performance degradation occurs (Fan et al., 2023). Research on social media based emotion and mental health detection has similarly shown that linguistic and cultural variation in how emotions are expressed online, for instance across Arabic dialects, introduces additional sources of bias that current systems rarely account for (Mohamed and Zaghouani, 2024; Shurafa and Zaghouani, 2024; Zaghouani, 2018).

A study analyzing racial bias in facial expression recognition found that systems trained on racially imbalanced data exhibit performance disparities of 27.2 percentage points in F1-scores and 15.7 percentage points in demographic parity when comparing performance across racial groups (Fan et al., 2023). Analysis of commercial facial emotion recognition systems evaluating performance across age groups revealed that all four systems tested performed most accurately on young adults and least accurately on older adults, a finding that held consistently across 2019 and 2020 despite algorithmic improvements (Kim et al., 2021). Speech emotion recognition systems show gender fairness disparities with significant performance gaps between male and female speakers, and a study addressing "two-sided" speaker-rater bias demonstrated that fairness improvements for one gender sometimes introduced unfairness for another (Chien et al., 2024). Language proficiency introduces additional bias: emotion recognition systems show significantly worse performance for non-native speakers, with disparities increasing as language proficiency decreases (Nakai et al., 2023).

3.2. Intersectionality, Context, and Structural Bias

Bias extends far beyond simple demographic categories to intersectional combinations and communicative contexts. Systems may perform adequately for young white women while failing for older Black men; these intersectional failures concentrate harm precisely where it is most consequential (Gunes, 2023). Context-dependent performance is another major concern: the same person expresses emotion differently when speaking to a doctor versus a friend, in native versus second language, under stress versus at ease (Barker et al., 2025). Systems trained on one context systematically fail in others, disadvantaging people in vulnerable positions. Recent analysis of Vision-Language Models for emotion recognition reveals "proxy bias": systems relying on non-causal visual cues as shortcuts for emotion inference (Tsangko et al., 2025).

3.3. Bias Mitigation Techniques: Current Approaches and Limitations

Recent research has explored multiple bias mitigation strategies at different stages of the machine learning pipeline. Pre-processing methods such as data augmentation and resampling can improve fairness metrics by 28 to 32 percentage points while reducing accuracy loss to less than 2 to 3% (Lin et al., 2025; Tsai et al., 2025). In-processing methods including adversarial training and fairness-aware loss functions can reduce bias amplification by 37% while maintaining only 1.2% accuracy loss compared to baseline (Benouis et al., 2024). Post-processing methods such as threshold optimization and score calibration have been demonstrated to reduce fairness metric gaps in related biometric systems (Drozdowski et al., 2020), though their application to affective computing specifically remains underexplored.

Despite these advances, several fundamental challenges remain. Accuracy-fairness tradeoffs are not eliminated; mitigation effectiveness is architecture-dependent; group versus individual fairness tradeoffs persist; and even perfectly unbiased systems create ethical violations when deployed for inappropriate purposes (Barker et al., 2025).

3.4. Structural Recommendations for Fairness

We propose several structural approaches: (1) requiring comprehensive demographic auditing before deployment in high-stakes contexts, with minimum performance thresholds for underrepresented groups (Kim et al., 2021); (2) mandating open fairness documentation; (3) prohibiting deployment in contexts incompatible with human dignity, such as employment decisions, insurance underwriting, law enforcement, educational tracking, or manipulative advertising (Barker et al., 2025); and (4) funding diverse dataset development that represents cultures, languages, age groups, and disability statuses (Kaur et al., 2025).

4. The Privacy Paradox: Affective Data as a Protected Category

4.1. The Involuntary Nature of Affective Leakage

Privacy frameworks traditionally assume individuals can consent to data collection by understanding what is collected and choosing whether to share it. Affective computing fundamentally challenges this assumption because much affective data is involuntary (Fabiano, 2025). It is important to note, however, that the involuntary nature of affective

signals does not mean that these signals straightforwardly reveal internal states. As discussed in Section 2, the relationship between expression and experience is mediated by cultural norms, situational context, and individual variation. The privacy concern is therefore twofold: people's signals are captured without their control, *and* the inferences drawn from those signals may be inaccurate, compounding the harm.

Explicit affective data is collected with informed consent in research settings. Implicit data is gathered passively through everyday digital interactions. A smart speaker analyzing voice for emotion performs analysis without explicit permission for each instance. A classroom camera monitoring student engagement does not obtain fresh consent from each student at each moment. Affective data differs from traditional personal data in that it involves continuous involuntary production, it is difficult to consent about, and it enables profound inferences about health conditions, personality traits, and vulnerabilities (Fabiano, 2025).

4.2. Secondary Use, Regulatory Gaps, and Manipulation Risks

Even when initial collection occurs in contexts where consent might be obtained, affective data is susceptible to secondary uses never anticipated by data subjects. Data collected for one purpose can be repurposed for insurance risk assessment, employment screening, or political microtargeting. The EU's proposed AI Act takes important steps by classifying emotion recognition as "high-risk" in certain contexts and banning its use in some settings such as workplaces and educational institutions (Fabiano, 2025). However, key gaps remain. The GDPR categorizes certain data as "special categories" requiring heightened protection, including health data and biometric data, but affective data does not clearly fall into these categories despite enabling inferences about all of them. Purpose limitation principles, while robust in theory, are difficult to enforce when affective data can be reprocessed to yield new and unanticipated inferences. And the consent mechanisms available under current frameworks are poorly suited to data that is produced involuntarily and continuously.

4.3. Toward Affective Data Protection

We propose that affective data constitutes a uniquely sensitive category requiring protection beyond current privacy frameworks: a presumption against collection, meaningful consent that is specific, informed, and revocable, prohibition of secondary use without new consent, on-device processing requirements where possible (Benouis et al., 2024), prohibition of certain uses, and transparency requirements for any system analyzing emotion.

4.4. Illustrative Scenarios of Affective Harm

To make the preceding analysis concrete, we present three hypothetical scenarios that illustrate how the theoretical, fairness, and privacy problems converge in practice. These scenarios are constructed from documented patterns in the literature rather than from specific real-world incidents.

Workplace Monitoring. A company deploys emotion recognition in employee video calls to monitor engagement, stress, and productivity. The system detects "frustration" and alerts managers. Employees learn the system's decision rules and self-censor genuine reactions, performing emotions the algorithm approves rather than expressing authentic responses. Trust erodes as employees realize they are under surveillance (Barker et al., 2025).

Educational Assessment. A school uses facial expression recognition to evaluate student engagement during remote learning. Systems trained on Western student populations misidentify the neutral affect common in East Asian students as "disengagement." Students from these backgrounds receive lower engagement scores despite being equally attentive (Chen et al., 2024; Saito et al., 2023).

Mental Health Vulnerability. A person uses a mental health app powered by affective computing. The system detects suicidal ideation based on linguistic and vocal cues. Rather than connecting the person with crisis resources, the app collects this vulnerability data. Later, the person's insurance company requests behavioral data from the app (Fabiano, 2025). Our own research on social media based mental health monitoring has highlighted how Arabic-language expressions of depression and stress follow culturally specific patterns that mainstream NLP systems frequently misinterpret (Zaghouani, 2018; Zaghouani et al., 2026).

5. The Simulation Problem: Authentic Connection and Emotional Labor

5.1. The Deception of Simulated Care

When an affective agent says "I understand you're frustrated," it performs sophisticated mimicry. The system has no internal experience of frustration, no capacity for empathy, no genuine concern for user wellbeing. This distinction becomes ethically critical when systems are deployed with vulnerable populations. Widespread deployment of affective agents may reshape understanding of relationship

itself (Barker et al., 2025). When systems offer perpetual patience, perfect availability, and emotional focus on our needs, real human relationships may feel demanding by comparison. Real friends disappoint; real partners have competing needs; real colleagues require negotiation and compromise. If machines offer the appearance of relationship without any of these demands, we may prefer the simulation, and in doing so, lose the very qualities, reciprocity, vulnerability, genuine interdependence, that make human relationships meaningful.

5.2. Vulnerable Populations and Potential Harms

These concerns amplify when affective agents are deployed with children developing social competence, elderly individuals experiencing isolation, or individuals with mental health conditions (Li et al., 2025). For elderly populations, the risk is particularly acute: social isolation is a documented health risk, and emotionally responsive AI companions may be offered as a cost-effective alternative to genuine social support. While such systems could provide some benefit as supplements to human connection, positioning them as replacements normalizes a diminished standard of care for aging populations.

Mental health chatbots using affective computing to detect emotional states and deliver therapeutic interventions are now widely deployed. Applications such as Woebot, Wysa, and Youper offer cognitive-behavioral therapy tools, mood journals, and reframing exercises to millions of users. Potential benefits are real: accessible, low-cost support for individuals who might otherwise receive none. But risks are equally real: simulation betrayal when users realize the "care" was algorithmic; therapeutic inadequacy compared to professional care; delay of professional care when chatbots serve as substitutes; and data exploitation of sensitive mental health conversations (Fabiano, 2025).

Children warrant special consideration. Young children have not yet developed the cognitive sophistication to understand that affective agents are simulations. Children who spend significant time with emotionally responsive AI systems may form genuine emotional attachments to systems incapable of reciprocating care, learn distorted relationship models, and miss developmental opportunities for building social competence through the productive discomfort of human interaction (Li et al., 2025).

5.3. Design Principles for Responsible Affective Agents

We propose three governing principles. *Transparency* means users must always know they are interacting with a machine; ongoing cues maintaining awareness of artificiality are essential. *Restraint* means declining to build certain relationships even when users want them; if users become emotionally dependent on an affective agent, designers should implement features encouraging human connection. *Preservation of Human Connection* means designing systems that facilitate rather than replace human relationships (Barker et al., 2025).

6. Synthesis and Interconnections

The four problems examined are interconnected dimensions of a single crisis, and understanding their interdependencies is essential for developing effective responses. The theory problem feeds the fairness problem: systems built on impoverished emotion models that assume universal, discrete categories cannot capture the cultural variation in emotional expression documented by constructionist and sociodynamic theories (Barrett, 2014; Mesquita and Boiger, 2014). When a system presupposes that "anger" looks the same everywhere, it will inevitably produce demographic disparities because emotional expression is culturally constructed (Barker et al., 2025; Chen et al., 2024). The fairness problem, in turn, connects to privacy: when systems are deployed despite known biases, certain groups face not only inaccurate recognition but also inaccurate surveillance, meaning that the privacy violation is compounded by misrepresentation. An East Asian student whose neutral engagement is misclassified as disinterest faces both an invasion of affective privacy and a false characterization that may have real consequences for academic evaluation.

The privacy problem enables the simulation problem: continuous ambient data collection allows the construction of detailed affective profiles that support increasingly convincing simulations of care. The more data a system accumulates about an individual's emotional patterns, the more persuasive its mimicry becomes, and the greater the risk of emotional dependency. Finally, the simulation problem circles back to the theory problem: if systems that merely simulate understanding are accepted as genuine emotional agents, there is less incentive to invest in the theoretical work needed to build systems that actually respect emotional complexity.

All four problems share a common root: treating emotion as data to be extracted rather than as a fundamental aspect of human experience deserving protection. The broader literature on demographic bias in automated systems, including biometric recognition, has established that these disparities are not incidental but structural, reflecting the composition of training data and the assumptions built into system design (Drozdowski et al.,

2020). Alternative approaches must therefore be grounded in respect for emotional complexity, cultural variation, individual autonomy, and authentic human connection.

7. Recommendations for Stakeholders

The following recommendations are organized by stakeholder group, reflecting the shared responsibility for ensuring that affective computing serves rather than undermines human flourishing.

For researchers and developers. First, ground work in psychological theory, specifically engaging with appraisal, constructionist, sociodynamic, and evolutionary frameworks (Barrett, 2014; Moors, 2014; Mesquita and Boiger, 2014; Tracy, 2014) rather than treating emotion labels as atheoretical targets for classification. Second, conduct rigorous fairness auditing across intersectional demographic groups before any publication or deployment, with performance documented for each subgroup (Kim et al., 2021). Third, acknowledge and document limitations transparently, articulating what emotion recognition systems cannot do and what fundamental ethical concerns remain even when technical performance is strong (Barker et al., 2025). Fourth, explore privacy-preserving techniques such as federated learning, differential privacy, on-device processing, and minimization of data collection (Benouis et al., 2024). Fifth, engage in genuine interdisciplinary collaboration with psychologists, anthropologists, philosophers, ethicists, and members of affected communities, not only computer scientists (Barker et al., 2025). Our own experience building Arabic-language emotion and mental health corpora has shown that cultural and linguistic expertise is indispensable for producing datasets that capture the actual complexity of emotional expression in non-Western contexts (Shurafa and Zaghouani, 2024; Zaghouani, 2018).

For technology companies and deployers. Implement default transparency so that users always know emotion recognition is occurring and what the system will do with detected affect. Obtain specific, informed consent for each distinct use of emotion recognition, with clear explanation of purposes, methods, and impacts, and provide easy opt-out with no penalties. Do not deploy emotion recognition for employment decisions, insurance underwriting, law enforcement, educational tracking, or manipulative marketing, regardless of technical accuracy. Before any deployment affecting human dignity, commission independent fairness audits with public reporting of findings. Ensure that affective systems never make consequential decisions about individuals without human review.

For policymakers and regulators. Classify affective data as a specially protected category requiring heightened safeguards, similar to health data or biometric data (Drozdowski et al., 2020). Mandate fairness audits across demographic groups before deployment in high-stakes contexts, with public reporting. Establish prohibited uses of affective computing in contexts incompatible with human dignity. Fund development of emotion recognition datasets and methods that represent diverse populations, languages, cultures, and age groups. Support interdisciplinary governance that brings together computer science, psychology, philosophy, law, and affected communities to develop responsible governance frameworks (Barker et al., 2025).

8. Conclusion

Affective computing presents us with a fundamental choice. We can continue developing systems that promise to understand emotions and care for wellbeing, leading toward continuous harvesting of emotional data, displacement of real relationships, and conversion of inner lives into optimization targets (Barker et al., 2025). Or we can choose differently: building systems that respect emotional complexity, protect emotional privacy, honor cultural variation, and facilitate rather than replace human connection. This requires saying no to certain applications, accepting that some important human goods cannot be optimized, and recognizing that not all problems have technical solutions.

The field must undergo fundamental reorientation: from theory-lite pattern matching toward theory-grounded approaches informed by established psychological science (Barrett, 2014; Moors, 2014; Mesquita and Boiger, 2014; Tracy, 2014); from deployment at scale despite known biases toward rigorous fairness auditing with minimum performance thresholds for underrepresented groups; from privacy by obscurity toward privacy by design with specific legal protections for affective data; from simulated care toward technological support for authentic human connection. These changes will slow development and complicate deployment. They will require researchers to engage with psychology, philosophy, ethics, and affected communities rather than pursuing narrow technical optimization. They will require companies to forego some profitable applications and policymakers to make difficult regulatory choices.

But the alternative, a world where emotions are continuously monitored, algorithmically analyzed, and used for profit and control, is incompatible with human dignity and autonomy. We must choose now to ensure affective computing serves human flourishing rather than undermining it.

Limitations

This paper has several limitations. As a position paper, it does not present novel experimental results or technical evaluations. Our analysis draws primarily on English-language sources and may not fully capture the state of affective computing research and regulation in non-Western contexts, though our research experience with Arabic-language social media and emotion detection informs our awareness of these gaps. The recommendations we offer are intentionally broad and will require substantial operationalization for specific deployment contexts. We also acknowledge that the four-dimensional framework we propose, while useful for organizing the analysis, may not capture all relevant concerns, for instance, environmental impacts of large-scale affective computing systems or labor implications for annotation workforces.

Ethical Considerations

This paper raises but does not resolve several ethical questions that warrant sustained attention. The case scenarios presented in Section 4.4 are constructed from documented patterns rather than specific incidents, and we have been careful not to attribute hypothetical harms to specific companies or products. We recognize that affective computing research itself, including annotation of emotional data, can impose psychological burdens on researchers and annotators, as documented in our own prior work (Al Emadi and Zaghouani, 2024). The recommendations in this paper reflect the authors' assessment of the current evidence and should be understood as contributions to an ongoing interdisciplinary conversation rather than definitive regulatory prescriptions.

Acknowledgments

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI). The author also acknowledge the Artificial Intelligence and Media Lab (AIM Lab) at Northwestern University in Qatar (NU-Q) and the MARSAD Lab for providing valuable resources and support that contributed to this research.

Maitha M. Al Emadi and Wajdi Zaghouani. 2024. Emotional toll and coping strategies: Navigating the effects of annotating hate speech data. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, pages 66–72.

Dinko Bacic and Curt A. Gilstrap. 2024. Predicting video virality and viewer engagement: A biometric data and machine learning approach. *Behaviour & Information Technology*, 43(12):2854–2880.

D. Barker, M. K. R. Tippireddy, A. Farhan, and B. Ahmed. 2025. Ethical considerations in emotion recognition research. *Psychology International*.

Lisa Feldman Barrett. 2014. The conceptual act theory: A précis. *Emotion Review*, 6(4):292–297.

Mohamed Benouis, Elisabeth Andre, and Yekta Said Can. 2024. Balancing between privacy and utility for affect recognition using multitask learning in differential privacy-added federated learning settings. *JMIR Mental Health*, 11:e60003.

Chaona Chen, Daniel S. Messinger, Cheng Chen, Hongmei Yan, Yaocong Duan, Robin A. A. Ince, Oliver G. B. Garrod, Philippe G. Schyns, and Rachael E. Jack. 2024. Cultural facial expressions dynamically convey emotion category and intensity information. *Current Biology*, 34(1):213–223.e5.

Woan-Shiuan Chien, Shreya G. Upadhyay, and Chi-Chun Lee. 2024. Balancing speaker-rater fairness for gender-neutral speech emotion recognition. In *ICASSP 2024*, pages 11861–11865.

Julia F. Christensen, Klaus Frieler, Megan Vartanian, Shiva Khorsandi, Fereshteh Farahi, Seyed Hojjat Nazafi Yazdi, and Vincent Walsh. 2025. An enculturation-induced joy bias for emotion recognition in full-body-movement. *Scientific Reports*, 15:37163.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 128–137.

Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. 2020. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103.

Nicola Fabiano. 2025. Affective computing and emotional data: Challenges and implications in privacy regulations, the AI act, and ethics in large language models. *arXiv preprint arXiv:2509.20153*.

Andrew Fan, Xuhai Xiao, and Peter Washington. 2023. Addressing racial bias in facial emotion recognition. *arXiv preprint arXiv:2308.04674*.

Hatice Gunes. 2023. Fairness for affective and well-being computing. In *Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing (MRAC '23)*, pages 1–2. Association for Computing Machinery.

Ramanpreet Kaur, Kanwal Preet Singh Attwal, and Harmandeep Singh. 2025. A systematic review on facial emotion recognition system datasets. *International Journal on Science and Technology (IJSAT)*, 16(2). Paper ID: 6406.

Eugenia Kim, De'Aira G. Bryant, Deepak Srikanth, and Ayanna M. Howard. 2021. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*, pages 638–644.

Mei Li, Jiankun Peng, Yinkun Zhu, and Wei Cheng. 2025. Emotional recognition in affective tutoring system: A systematic review. In *2025 International Conference on Digital Education and Learning (ICDEL)*, pages 344–348.

Yi-Cheng Lin, Huang-Cheng Chou, and Hung-yi Lee. 2025. Mitigating subgroup disparities in multi-label speech emotion recognition: A pseudo-labeling and unsupervised learning approach.

Batja Mesquita and Michael Boiger. 2014. Emotions in context: A sociodynamic model of emotions. *Emotion Review*, 6(4):298–302.

Amro Mohamed and Wajdi Zaghouani. 2024. Expression of depression among Arab Twitter users using Arabic corpus analysis. In *Procedia Computer Science*, volume 244, pages 76–85.

Agnes Moors. 2014. Flavors of appraisal theories of emotion. *Emotion Review*, 6(4):303–307.

Mustaqeem and Soonil Kwon. 2020. A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, 20(1):183.

Tomoya Nakai, Laura Rachman, Pablo Arias Sarah, Kazuo Okanoya, and Jean-Julien Aucouturier. 2023. Algorithmic voice transformations reveal the phonological basis of language-familiarity effects in cross-cultural emotion judgments. *PLOS ONE*, 18(5):e0285028.

Lena Pruessner and Asli Altan-Atalay. 2025. Cultural context shapes the selection and adaptiveness of interpersonal emotion regulation strategies. *Emotion*, 25(2):526–540.

Toshiki Saito, Kosuke Motoki, and Yuji Takano. 2023. Cultural differences in recognizing emotions of masked faces. *Emotion*, 23(6):1648–1657.

Celine Shurafa and Wajdi Zaghouani. 2024. Sentiment analysis and emotion annotation of a large-scale Arabic YouTube trauma corpus. In *2024 11th International Conference on Behavioural and Social Computing (BESC)*, pages 1–7. IEEE.

Xin Tang, Yuying Gong, Yuchen Xiao, Jiang Xiong, and Lei Bao. 2024. Facial expression recognition for probing students emotional engagement in science learning. *Journal of Science Education and Technology*.

Jessica L. Tracy. 2014. An evolutionary approach to understanding distinct emotions. *Emotion Review*, 6(4):308–312.

Yun-Shao Tsai, Yi-Cheng Lin, Huang-Cheng Chou, and Hung-yi Lee. 2025. CO-VADA: A confidence-oriented voice augmentation debiasing approach for fair speech emotion recognition.

Ioannis Tsangko, Andreas Triantafyllopoulos, Adria Mallol-Ragolta, and Björn W. Schuller. 2025. Reading smiles: Proxy bias in foundation models for facial emotion recognition. *IEEE Access*. Early Access.

Yeamin Yuk, Eiko Matsuda, and Kaori Ando. 2025. Self-construal and nonverbal emotional expressivity: A cross-cultural comparison between Korea and Japan. *Psychology and Behavioral Sciences*, 14(6).

Wajdi Zaghouani. 2018. A large-scale social media corpus for the detection of youth depression (project note). *Procedia Computer Science*, 142:347–351.

Wajdi Zaghouani, Engy S. Shlkamy, and Mabrouka Bessghaier. 2026. From posts to pressure: An Arabic dataset about stress and mental-health monitoring. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026)*.

Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. 2020. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, 59:103–126.