EDITOR: Erik Cambria, cambria@ntu.edu.sg

DEPARTMENT: AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS

The Emotion Labeling Problem in Affective Computing Research

Christian Montag, Chengzhong Xu, Michiel Spapé, *University of Macau, China*

Erik Cambria, *Nanyang Technological University, Singapore*

Abstract—Artificial intelligence (AI) has predominantly focused on replicating human cognitive abilities, yet for AI to interact meaningfully with users it must also exhibit emotional intelligence. To enable this, affective computing must allow accurately sensing human emotions such as through facial expressions, but due to the subjective nature of emotions, this approach is complicated by what we term the Emotion Labeling Problem. While the long-standing basic emotion theory posits that emotions like joy and anger are universally expressed and understood, social constructivists challenge this claim, arguing that emotional interpretation is highly context-dependent. Mislabeling emotions in AI datasets leads to invalid emotional recognition, for example when facial expressions do not align with true internal states. As potential solutions, we propose self-reported emotion labeling, behavioral description approaches, leveraging metaphor understanding for nuanced emotional inference, and modeling emotional triggers by means of causal emotion entailment. Addressing these labeling challenges is critical to improving the capability of AI to detect and respond proactively to human emotions, facilitating more empathetic and psychologically informed human-AI interaction.

Since its inception at the Dartmouth Summer Research Project in 1956 [1], the field of AI has primarily concentrated on replicating human cognitive abilities. However, it has become increasingly evident that effective human-computer interaction also depends on AI systems that can respond with emotional warmth and empathy [2]. This is especially important in healthcare settings. To enable emotionally intelligent AI, a key requirement is the accurate sensing of the affective states of individuals interacting with the system [3]. Different human modalities can be used to sense emotions including the prosody of language, the gait of a person and of course facial expressions [4]. In this realm, it is of interest that the computer sciences rely until today strongly on the concept of basic emotions to make sense of human facial expressions.

1541-1672 © 2025 IEEE
Digital Object Identifier 10.1109/MIS.2025.XXXXXXX

This theory goes back to Charles Darwin's "The expressions of emotions in man and animals" [5] and has been championed in the last 50 years by Paul Ekman's work [6]. A key premise of basic emotion theory is that a common set of emotions such as joy, anger and sadness are universally shown in our species and of note also universally understood when humans interact with each other [7]. Although this theory is backed up by some evidence, it has also been challenged by social constructivists, whose experiments showed that the naming of emotional expressions is not as cross-culturally consistent, and therefore not universal, as basic emotion theorists argued [8]. Further, an understanding of emotional expressions needs always also to be seen in the context of a situation. These factors are critical for affective computing: even if a machine 'correctly' detect a happy facial expression, it does not mean the person actually felt happy. Perhaps he or she was in deep mourning, but hid their sadness by cognitive control.

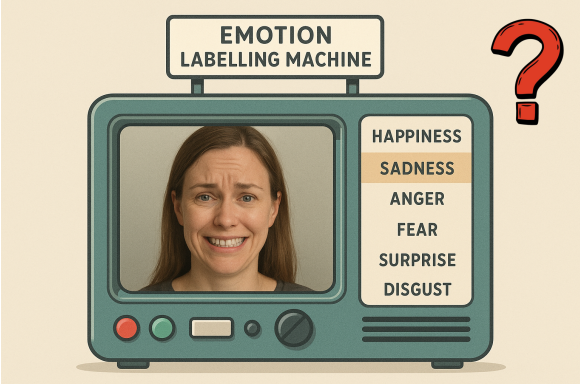

FIGURE 1. The *Emotion Labeling Problem* refers to AI mislabeling emotions, like when facial expressions don't match actual feelings, resulting in flawed emotion recognition.

Against this background, we argue that research in affective computing is hampered by what we call the *Emotion Labeling Problem* of datasets which will are needed for training AI systems in the context of affective computing. Problems with labeling might arise i) if someone feels and expresses emotion A (e.g., disgust), but it is labeled B (e.g., anger) and ii) if someone feels emotion joy and expresses some mixture of joy and neutral expression, and it is labeled joy. Also other problems arise when someone feel emotion A and expresses a mixture of A and B, with the outcome of labeling emotion C. And of course there can random annotation errors. In an opinion piece [9] on a Nature study inferring emotions from YouTube videos via deep neural networks it was put forward [10] that a solution might be to do labeling more in behavioral ways - such as speaking of wide-eyed faces - to not fall into the trap of labeling emotional states which might not be felt.

Doing this might not represent the full solution to the problem because somehow emotional states of a person depicted in pictures, videos or else must be labeled regarding emotions – otherwise the machine will not learn about an emotional state, but just about a behavior of a person. What kind of other solutions could there be? We suggest that one avenue to go would be to do instant labeling of a video, voice or picture by the person who is depicted. This way we would not need to rely on labeling from the outside (hence doing mind-reading from faces). This approach of course also has disadvantages, namely if the person has insufficient introspective skills. Nevertheless, it would be interesting to systematically compare how well emotional expressions can be depicted from different modalities of humans when comparing the labeling done via the different described avenues here. Further, we need also to study the brain mechanisms underlying emotions and facial expressions and only by mapping both the inner and out view onto each other, we will get closer to an understanding of what a person felt.

Another potential solution involves leveraging metaphor understanding to gain deeper insights into how users truly feel about specific topics. By analyzing the metaphors people use in their language, we can identify underlying emotional and conceptual associations. These metaphors often reflect abstract feelings and attitudes that may not be directly expressed. Through computational modeling of concept mappings derived from these metaphorical expressions, AI systems can infer users' emotional states and perspectives with greater nuance and accuracy [11]. For this method to be effective, however, it requires not only a large volume of user text but also a consistent and meaningful use of metaphorical expressions. On a related note, having access to large volumes of user-generated text is crucial for effective emotion detection, as emotional expressions are highly context-dependent. The same phrase or piece of text can convey different emotional meanings depending on the user's personality traits, individual preferences, cultural background, and personal experiences [12]. Without sufficient textual data to model these user-specific factors, AI systems risk misinterpreting emotional cues, leading to inaccurate or overly generalized assessments. Therefore, personalized emotion recognition requires not only sophisticated models but also rich, contextually grounded language data.

Another promising approach to preventing – or at least mitigating – the *Emotion Labeling Problem* is the use of causal emotion entailment [13]. This method involves identifying and modeling the causal relationships between events, user experiences, and the emotional reactions they elicit. Rather than relying solely on surface-level emotion annotations, causal emotion entailment seeks to understand the underlying reasons why a particular emotional response occurs in a given context. By mapping these cause-and-effect chains, AI systems can infer emotions more accurately, even in cases where explicit labels are missing, ambiguous, or inconsistent. This approach not only enhances the reliability of emotion recognition models but also aligns more closely with how emotions are understood in psychological and cognitive sciences, where emotions are often seen as responses to goal-relevant appraisals and situational triggers.

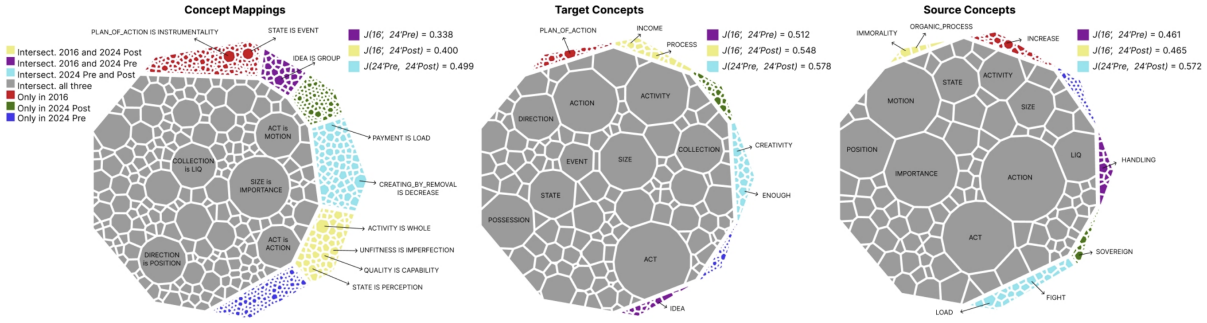

FIGURE 2. Understanding metaphors helps AI uncover deeper emotional insights by revealing abstract feelings and attitudes not directly stated. Modeling these concept mappings allows more nuanced and accurate emotion detection.

The *Emotion Labeling Problem* has critical implications beyond emotional theories and concerns of reliability in affective computing. A key area of Emotion AI and affective computing deals not only with the prediction of emotional states of a person, but even with the prediction of affective disorders such as depression. Here, also problems with labeling can appear, if - the prediction of two groups will be done, let's say depressed vs. healthy controls. Here, the problem lies on the one hand on the fact that a depressed states comes in many variations (e.g., major depression, dysthymia or psychotic depression). Hence lumping them all together into one category will blur the results which can be achieved when not doing more fine granular labeling. On the other hand, it has been questioned already by former NIMH director Tom Insel if the current categorization system of mental disorders via DSM-5 (then version 4) actually works. Instead Insel proposed the research domain criteria (RDoC) for the study of mental disorders [14], whereas the focus in the study of mental disorders was shifted for instance from DSM-diagnoses to smaller units of analysis including layers of genetics, brain imaging, behavior and so forth. By focusing on the prediction of subunits underlying mental disorders - such as altered emotional processes in the brain - we might come closer to an understanding of psychopathologies. Following such an approach in the study of emotional AI might reduce noise in analyzed datasets. Instead of labeling a depressed vs. non-depressed mind, it would be then better to directly predict altered brain mechanisms, which likely also overlap in several brain disorders (this also explains why the current categorization-system is flawed). But this of course would require to not predict from emotional faces depression, but from emotional faces certain brain mechanisms or the other way round. This is also in line with the idea that there is a view from outside on emotional expressions (studying faces) and from the inside on emotional expressions (studying brain mechanisms [15]).

These brief examples highlight the critical importance of addressing the *Emotion Labeling Problem* in order to advance the development of Emotion AI systems. Accurately labeling emotional data remains one of the most significant challenges, as it directly affects the ability of AI to recognize, interpret, and respond to human emotions in meaningful ways. Just think of a person which is actually sad, but due to societal constraints actually wants to appear to be happy. A potent AI system would at best be able to flag this.

The examples put forward in this short article underscore the value of integrating insights from psychology into computer science—particularly the ongoing debates between basic emotion theory and social constructivist perspectives. Acknowledging and engaging with these theoretical frameworks allows researchers and engineers to build models that are not only technically robust but also aligned with the complexity and variability of human emotional expression. Bridging these disciplinary perspectives is therefore an essential step toward developing AI systems that can interact with users in a more nuanced, empathetic, and human-centered manner.

REFERENCES

1. J. McCarthy et al., "A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955," *AI magazine*, vol. 27, no. 4, 2006, pp. 12–12.
2. H. Zhang et al., "Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark," *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 2872–2881.

3. Q. Lin et al., “Has multimodal learning delivered universal intelligence in healthcare? A comprehensive survey,” *Information Fusion*, 2024, p. 102795.
4. G. Hu et al., “Recent trends of multimodal affective computing: A survey from NLP perspective,” *arXiv preprint arXiv:2409.07388*, 2024.
5. C. Darwin, *The expression of the emotions in man and animals*, anboco, 2016.
6. P. Ekman, “An argument for basic emotions,” *Cognition & emotion*, vol. 6, no. 3-4, 1992, pp. 169–200.
7. Z. Wang, S.-B. Ho, and E. Cambria, “A review of emotion sensing: categorization models and algorithms,” *Multimedia Tools and Applications*, vol. 79, 2020, pp. 35553–35582.
8. W. D. TenHouten, “Basic emotion theory, social constructionism, and the universal ethogram,” *Social Science Information*, vol. 60, no. 4, 2021, pp. 610–630.
9. L. Barrett, “Debate about universal facial expressions goes big,” *Nature*, vol. 589, 2021, pp. 202–203.
10. A. S. Cowen et al., “Sixteen facial expressions occur in similar contexts worldwide,” *Nature*, vol. 589, no. 7841, 2021, pp. 251–257.
11. R. Mao et al., “Unveiling diplomatic narratives: Analyzing United Nations Security Council debates through metaphorical cognition,” *Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 46*, 2024.
12. L. Zhu et al., “Neurosymbolic AI for personalized sentiment analysis,” *International Conference on Human-Computer Interaction*, 2024, pp. 269–290.
13. H. Liu et al., “Knowing What and Why: Causal emotion entailment for emotion recognition in conversations,” *Expert Systems With Applications*, vol. 274, 2025, p. 126924.
14. T. Insel et al., “Research domain criteria (RDoC): toward a new classification framework for research on mental disorders,” 2010.
15. C. Montag and J. Panksepp, “Primal emotional-affective expressive foundations of human facial expression,” *Motivation and Emotion*, vol. 40, 2016, pp. 760–766.

Christian Montag is currently working as Distinguished Professor of Cognitive and Brain Sciences at the Institute of Collaborative Innovation with the University of Macau, Macau SAR, China. His research interests include digital phenotyping, behavioral addictions and how AI impacts upon societies. Montag received a PhD in Psychology from University of Bonn, Bonn, Germany. Contact him at cmontag@um.edu.mo

Chengzhong Xu is currently a Chair Professor of Computer Science at the Faculty of Science and Technology at University of Macau, Macau SAR, China. His research interests lie in Parallel and Distributed Computing, Cloud and Edge Systems for AI, Intelligent Transportation and Autonomous Driving and Reinforcement Learning and AI Theory. Xu was awarded a PhD in Computer Science from Hong Kong University in Hong Kong SAR, China. Contact him at czxu@um.edu.mo.

Michiel Spapé is Associate Professor in Cognitive Neuroscience at the Centre for Cognitive and Brain Sciences, University of Macau, Macau SAR. His research interests involve cognition and consciousness, affective neuroscience, and neuroaffective interaction. Spapé received his PhD in Psychology from Leiden University, Netherlands. Contact him at mspape@um.edu.mo

Erik Cambria is currently a Professor of Artificial Intelligence at NTU CCDS. His research focuses on neurosymbolic AI for interpretable, trustworthy, and explainable affective computing in domains like mental health, climate resilience, and socially responsible investing. Contact him at cambria@ntu.edu.sg.