Towards Context-sensitive Emotion Recognition

Mukherjee, Sayak

DOI
10.1145/3716553.3750824

Publication date
2025

Document Version
Final published version

Published in
ICMI 2025 - Proceedings of the 27th International Conference on Multimodal Interaction

Towards Context-sensitive Emotion Recognition

Sayak Mukherjee
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Delft, South Holland, Netherlands
s.mukherjee-3@tudelft.nl

Abstract

Achieving socially compatible human-AI interaction requires systems that can interpret and respond to human emotions appropriately in complex social environments. While traditional emotion recognition models rely heavily on facial or bodily expressions, a growing body of research demonstrates that such cues are insufficient without the dynamic, multimodal contextual cues. Positioned at the intersection of cognitive psychology and AI, this work identifies three essential qualities for context-sensitive emotion recognition (CSER): generalizability to unseen scenarios, data efficiency in adapting to new contexts, and reliability in predictive performance across contexts. We outline a research plan that systematically investigates the role of contextual factors, domain adaptation, and uncertainty quantification in building CSER models capable of robust performance across real-world settings. Our approach integrates computational rigour with ethical responsibility to lay the foundation for next-generation emotion-aware systems that are not only accurate but also trustworthy, transparent, and support human well-being in digital interactions.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Cognitive science**; **Theory of mind**.

Keywords

Context-aware Emotion Recognition, Affective Computing, Context, Data-efficiency, Generalisation

ACM Reference Format:
Sayak Mukherjee. 2025. Towards Context-sensitive Emotion Recognition. In *Proceedings of the 27th International Conference on Multimodal Interaction (ICMI '25), October 13–17, 2025, Canberra, ACT, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3716553.3750824

1 Introduction

Emotion is fundamental to human social cognition, shaping how we interpret others' intentions, beliefs, and values [20]. Extensive evidence suggests that emotions modulate key cognitive functions [45], including perception, learning, memory, reasoning, and problem-solving, by prioritising emotionally salient stimuli [41]. Thus, in a world where we envision intelligent systems closely collaborating with humans in complex environments, such as healthcare or education, the ability to automatically infer emotions is essential [31, 34]. It not only allows for modelling human experience but also for capturing human utility functions for the development of *human-compatible* [38] intelligent systems.

Despite its critical role in social interaction, emotion recognition in computational models has traditionally been framed as a supervised pattern recognition problem [48], primarily based on facial or bodily expressions. However, growing empirical evidence indicates that expressions alone are insufficient for accurate affective inference, as human emotion perceptions are deeply influenced by contextual factors [3, 10, 37]. In response, the field of affective computing has started to explore context-aware emotion recognition (CAER), with datasets and models that explicitly incorporate contextual information [24]. Integrating such features has been shown to significantly enhance model performance, particularly by accounting for subjectivity and ambiguity inherent in emotional expression [29]. Nevertheless, CAER presents substantial challenges due to its inherently multimodal, dynamic and temporally extended nature [4]. Effective emotion recognition often depends on a complex interplay of facial cues, speech, prosody, body posture, environmental cues, and social dynamics. These components span diverse data modalities, including text, audio, video, and images, demanding sophisticated multimodal integration strategies for robust modelling.

Many existing CAER approaches focus on minimising predictive error within curated, labelled datasets [9, 14, 16]. However, for intelligent agents to collaborate effectively in complex, real-world environments, they must generalise beyond narrowly defined training distributions. This generalisation remains a key challenge due to the inherent context-sensitive nature of emotion recognition, where expressions and emotions are intricately shaped by contextual cues. Capturing the breadth of possible relationships between expression, context, and emotion labels would require large-scale, diverse data collection across numerous environments. Yet, such data acquisition is often constrained by privacy, logistical, and ethical concerns, particularly in sensitive domains such as healthcare and education. These challenges highlight the need to develop multimodal models for context-sensitive emotion recognition (CSER) that can operate robustly across diverse environments. In this work, we define CSER as an encompassing framework that includes both context-aware and context-invariant approaches, as well as ensembles thereof, all aimed at achieving reliable generalisation across diverse settings.

Emotion recognition, as a sensitive application, has recently come under increasing scrutiny from policymakers, particularly in light of regulations such as the GDPR and the EU AI Act. This regulatory attention has prompted a growing shift among researchers and developers toward ethically aligned and responsible design practices for emotion recognition technologies [30]. In this context, it is essential to recognise that building robust CSER models that

ICMI '25, Canberra, ACT, Australia

ACM ISBN 979-8-4007-1499-3/25/10
https://doi.org/10.1145/3716553.3750824

perform reliably across previously unseen settings remains a significant challenge. Thus, for high-stakes domains such as healthcare, where incorrect predictions can lead to harmful consequences, it is often preferable for a system to abstain from making a prediction when uncertainty exceeds acceptable thresholds rather than risk an erroneous one [19]. To support this endeavour, CSER models must be capable of quantifying uncertainty. While multimodal emotion recognition has attracted interest from both academia and industry, there remains limited focus on integrating uncertainty quantification into CSER systems. In this work, we argue that advancing computational models of emotion recognition requires not only improving their generalisability but also ensuring their reliability through principled uncertainty estimation.

In the following sections, we begin by discussing our research objective, Section 2. Then, we provide a brief review of the related works in Section 3. In Section 4, we describe our current progress, which aims to answer the first research question. Finally, Section 5 outlines our planned future work, which aims to answer the remaining questions.

2 Research Objective

Our goal is to develop a multimodal CSER model that is effective across diverse real-world settings. For this, we start by defining the *qualities of effectiveness* (QEs) for a multimodal CSER model as:

- **Generalisability**. The ability to perform well in previously unseen, novel contexts.
- **Data-efficiency**. The capacity to adapt to new contexts with minimal supervision or data [44, 47].
- **Reliability**. Consistent predictive performance across diverse conditions.

Towards achieving our goal, we identify three key research challenges. We summarise them below, along with the corresponding research questions aimed at addressing each challenge.

C1. *Effect of context in generalisability.* Emotion recognition is deeply influenced by a range of contextual cues, including personal, situational, and cognitive factors. As a result, the CAER models trained on specific, curated datasets would struggle to generalise to novel environments where one or more contextual dimensions differ, leading to distribution shifts. The nature and severity of this impact would vary depending on the type of distribution shift encountered, such as covariate shift, label shift, or concept shift, each of which may arise from changes in specific contextual variables between training and test conditions. Moreover, generalisation may be affected by spatial and temporal contexts [36]. Despite the clear dependence of emotion recognition on contextual factors, there is a lack of principled investigations into how different types of context affect model generalizability. To address this gap, we pose the following research questions:

- **RQ1a**. What contextual factors impact generalisation?
- **RQ1b**. How does each contextual factor affect a model's ability to generalise across different settings?

C2. *Data-scarcity in real-world contexts.* To enable CSER models to perform effectively in new, unseen environments, a straightforward solution might involve collecting large volumes of data from the target domain and fine-tuning the model by minimising supervised predictive error. However, this strategy is often impractical due to logistical constraints, privacy concerns, and ethical limitations associated with data collection in real-world settings. Consequently, there is a pressing need for CSER models that can adapt to novel domains using minimal data and supervision. Approaches such as meta-learning have shown promise in this regard, offering mechanisms for rapid adaptation with limited examples. Similarly, recent advances in causal representation learning [40] suggest it may enable more data-efficient generalisation across domains. Despite this, there is a notable lack of comparative studies evaluating the data efficiency of these adaptation methods specifically within the context of CSER. To address this research gap, we pose the following question:

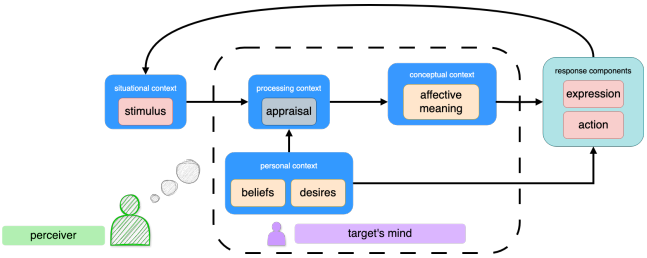

Figure 1: Planning over Emotional Theory of Mind. Style: Red indicates observable states, Orange represents unobservable states; Grey represents processing units. Blue represents the contextual factor for each component [9].

- **RQ2**. What is the data efficiency of different domain adaptation methods for CSER approaches?

C3. *Quantifying uncertainty towards reliable CSER.* Errors in emotion recognition can have serious consequences, and even psychological harm, particularly in high-stakes environments such as healthcare. To ensure ethically aligned design of CSER systems, it is therefore critical that these models can assess and communicate the risk associated with their predictions. A principled approach to uncertainty quantification would enable CSER models to abstain from making unreliable predictions, thereby enhancing their trustworthiness and safety. Despite its importance, uncertainty quantification has received limited attention in the development of CSER models. To address this gap, we pose the following research question.

- **RQ3**. How to quantify the uncertainty of CSER models?

3 Related Work

Emotions are structured psychological states [33] that arise in response to internal or external events that are appraised [39] as significant to an individual's goals, beliefs, and values. These states involve coordinated changes across experience, expression, physiology, and behaviour. Owing to recent developments in computational cognitive modelling, researchers have been able to test hypotheses about how people make inferences about the social world [1, 2]. Within this broader effort, our focus lies on *affective cognition* [31], which is the computational study of how individuals infer, reason about, and respond to the emotional states of others. In particular, we base our review of related work in the framework of Emotional

Theory of Mind (EToM) [20, 32], which characterises emotion inference in terms of two primary causal relationships: antecedents (i.e., "What does the perceiver believe caused the target's emotional state?") and descendants (i.e., "How does the perceiver believe the target's emotion influences their behaviour or expression?"), as illustrated in Figure 1. Throughout the discussion, we refer to the person making the inference as the *perceiver*, and the person whose emotional state is being inferred as the *target*. Using this framework, we organise the literature into two major categories of affective inference: (i) Emotion Recognition, which involves reasoning "backward" from the target's observable behaviour to their emotion, and (ii) Context-Aware Emotion Recognition, which extends emotion recognition by incorporating the "forward" reasoning about the target's emotion given contextual cues such as situational factors, stimuli, and the target's other mental states.

Emotion Recognition. Earlier works treated emotional expressions as biologically rooted and largely universal [11]. This perspective led to the development of computational models focused primarily on inferring emotion directly from facial expressions. These models typically relied on feature extraction methods such as the Facial Action Coding System (FACS) [12], followed by supervised learning techniques to map expressions to discrete emotion categories or continuous dimensions like Valence-Arousal-Dominance (VAD). This closely resembles Thinking Fast, or System 1 thinking, in the dual-process model of cognitive processes [23], which operates quickly and intuitively, based on certain heuristics. However, recent research has increasingly challenged this view. While emotions may manifest externally as facial or bodily expressions, these expressions do not carry intrinsic meaning [3, 4, 49]. Instead, affective cognition depends on the integration of expressive cues with contextual information and inferences about the mental states of the observed individual [8, 22]. For example, crying may express sadness, joy, relief, or frustration. The same facial configuration comprising tears and a contorted face could indicate vastly different emotional states depending on the situation (e.g., at a funeral versus a wedding). Moreover, individuals from different cultural backgrounds often vary in their perceptions of emotions [15]. These findings underscore the inherent ambiguity and context-dependence in mapping expressions to emotional states, motivating a shift toward context-aware approaches.

Context-Aware Emotion Recognition. Contrary to earlier works, recent research has underscored the critical role of context in shaping how emotions are perceived. Context refers to the collection of environmental, social, and cultural factors that provide interpretive scaffolding for emotional inference [5]. Notably, observers also bring prior knowledge into supposedly context-free settings, including task-specific expectations [6] and conceptual priors accumulated through lived experience [7]. To account for the joint influence of expression and context, recent approaches in CAER have incorporated both backwards reasoning from expression and forward inference from context. These models typically treat the extraction of emotional cues from expressions and context as distinct processes, which are later integrated into a unified decision framework. This process is referred to as *emotional cue integration* [32, 42]. A key challenge in this integration is extracting meaningful emotional cues from the context. One well-known framework guiding this effort is Appraisal Theory [35], which posits that emotions arise from how individuals appraise events relative to their goals, beliefs, and values. Computational models that implement this theory aim to perform *third-person appraisals* [21, 50, 51], where the perceiver infers how a target might appraise a given situation. Within a dual-process theory of cognition, this corresponds to System 2, or Thinking Slow, which involves deliberative, logical reasoning and causal inference. However, modelling third-person appraisal presents several challenges. First, defining a complete set of appraisal dimensions either requires adaptation from first-person appraisal theories, which would require empirical validation in third-person contexts, or relies on data-driven similarity analyses [43], which are computationally expensive and methodologically complex. Second, establishing a mapping from appraisal dimensions to emotion labels is non-trivial, often hindered by the limited availability of annotated data. Third, reliably extracting appraisal dimensions from real-world data remains an open challenge. To address these limitations, researchers have leveraged vision-based pipelines [29] and large language models (LLMs) [14] to extract contextual features from visual scenes [24, 26] or event descriptions [21], and have used end-to-end learning frameworks to infer emotion labels from these cues [13]. While they offer promise, there is a lack of structured analysis on their effectiveness across different contexts, as well as potential biases in closed-source models. Furthermore, it has been shown that fine-tuning these models on CAER datasets improves performance. However, this is constrained by the ability to collect relevant data across diverse real-world settings.

4 Ongoing Research

Currently, our focus is on answering **RQ1a** and **RQ1b**. While recent advances in end-to-end deep learning-based approaches leveraging VLMs and LLMs have shown promising results, they leave open a critical question: *Which contextual features are most important?* [18] This gap in understanding largely stems from the inherent difficulty of abstracting and reasoning over the complex, multidimensional nature of real-world contexts. As models grow in scale and complexity, we argue that systematically identifying and analysing the contribution of different contextual factors is not only overdue but essential for advancing CSER. First, such an understanding would guide more strategic and privacy-conscious data collection, especially important given the ethical and logistical constraints of gathering affective data in sensitive environments. Second, it would inform model design by enabling the selection of the most discriminative and relevant contextual features, leading to simpler, more efficient, and more interpretable models. Additionally, it would help us to identify unfair biases. Third, it would provide a principled basis for understanding and managing domain shifts by revealing contextual variations whose value, when changed, is likely to cause a domain shift. Addressing this gap is therefore a necessary step toward building CSER systems that are transparent and generalisable for real-world, high-stakes applications.

Datasets. For our analysis, we utilise the EMOTIC dataset [24], which comprises a large collection of images with visual context sourced from the web and annotated with 26 discrete emotional

labels. It is important to acknowledge the limitations of EMOTIC. Specifically, the labels are based on third-person annotations rather than self-reports, which restricts our study to examining perceived emotion rather than felt emotion. Furthermore, the dataset lacks metadata about the annotators, which limits our ability to explore how a perceiver's context influences their emotion judgments.

Contextual Features. To explore the role of context in emotion recognition, we focus on a selected subset of contextual factors related to both the target and the situation. In the preliminary phase of our study, we examine two target-related factors: age group and cultural background. These were chosen due to practical challenges associated with data availability for certain instances of these contextual factors. For instance, while data collection involving younger individuals is relatively feasible, acquiring representative data from elderly populations often encounters privacy, ethical, and logistical constraints. Similarly, cultural diversity in existing datasets is limited, with certain cultural groups significantly underrepresented. Collecting balanced, large-scale data across global cultures poses considerable difficulty, especially when aiming to capture authentic emotional expressions in natural settings. In terms of situational context, we analyse the social setting, distinguishing between formal and informal environments, and between individuals who are alone versus those in groups. These settings influence the expression and perception of emotion, yet capturing data in certain configurations (e.g., formal gatherings) remains challenging due to access limitations and ethical considerations.

A key challenge we encountered is the absence of annotations for these contextual dimensions in the EMOTIC dataset. One approach would involve manually annotating each image or using crowdsourcing to label contextual features. However, given the exploratory nature of our preliminary study and the need for a time- and cost-efficient method, we will employ a Vision-Language Model (VLM). The model will be prompted with each image and tasked with inferring contextual labels across the selected dimensions. This automated approach provides a scalable mechanism for estimating contextual metadata and serves as a first step toward identifying the most important contextual factors.

Proposed Methodology. To systematically evaluate the importance of individual contextual factors in affective perception, we adopt a domain shift–based framework. Our central hypothesis is that *a contextual factor is critical if altering its value induces a significant shift in the data distribution, thereby affecting the CAER model's performance.* To this end, we employ the depth- and GCN-based EMOTICON model [29], and the CAER-Net model [26]. For each contextual factor under investigation, we partition the EMOTIC dataset into subsets according to the factor's distinct values (e.g., different age groups or social settings).

We then assess domain shift along three complementary axes. First, we measure *cross-split generalisation performance*: a model is trained on one subset and evaluated on the others to observe the change in accuracy. Second, we estimate the *data efficiency required for adaptation* by computing sample-wise learning curves when fine-tuning a model trained on one subset with data from other subsets. Third, we quantify *distributional divergence* directly in the feature space of the CAER model using statistical measures such as Maximum Mean Discrepancy (MMD) [17] and sliced Wasserstein distance [46], computed over intermediate representations obtained from models trained on different subsets. Together, these analyses allow us to identify which contextual dimensions induce the greatest distributional shift and, consequently, are most influential in shaping emotion recognition performance.

5 Future Research Directions

We outline our future research goals to answer RQs 2 and 3.

Generalisability. Building on the contextual factors identified in Section 4, a key direction for future work is to develop strategies that enable data-efficient generalisation across domain shifts induced by changes in contextual variables. Traditional approaches, such as fine-tuning large pre-trained models or applying meta-learning, typically rely on either substantial adaptation data or prior exposure to all relevant contexts during training. These assumptions are often unrealistic in real-world scenarios, where AI systems are expected to function reliably with minimal or no data from unseen settings. To overcome these limitations, we hypothesise that shifting the focus from purely correlational modelling to learning causal representations [40] for emotion recognition can support the development of robust, transferable representations that generalise in a few-shot setting. Recent advances in causal representation learning have demonstrated promise in capturing the compositional [25, 28] and generalisable [27] structure of human reasoning, aligning closely with the cognitive mechanisms involved in emotion understanding [16]. This research question will investigate the validity of this hypothesis and systematically assess the trade-offs between data efficiency and predictive performance.

Reliability. Emotion recognition errors can lead to miscommunication, reduced user trust, or even psychological harm in sensitive applications. Therefore, CAER systems must be able to quantify the uncertainty in their predictions, particularly when faced with ambiguous or out-of-distribution inputs. This research question focuses on modelling both epistemic (model-based) and aleatoric (data-based) uncertainty in affective inference. The objective is to develop systems that can defer judgment when necessary to human experts, upholding safety, ethical standards, and human alignment.

6 Safe and Responsible Innovation Statement

Broader Impact. Identifying contextual factors in emotion recognition improves interpretability, enhances data efficiency, and reduces bias by revealing hidden distributional imbalances. By promoting generalisability, transparency, and fairness, our work advances the development of robust, trustworthy, and emotion-aware systems that support human well-being in digital interactions.

Ethical Considerations. Emotion recognition is classified as a "high-risk" application under the EU AI Act, particularly in sensitive domains. Our research not only addresses the generalisability of CSER models but also prioritises reliability through models that transparently quantify uncertainty. This supports regulatory compliance while reflecting a principled commitment to developing human-centred and socially responsible AI systems.

References

[1] Chris L Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour* 1, 4 (2017), 0064.

[2] Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. 2009. Action understanding as inverse planning. *Cognition* 113, 3 (2009), 329–349.

[3] Lisa Feldman Barrett. 2022. Context reconsidered: Complex signal ensembles, relational meaning, and population thinking in psychological science. *American Psychologist* 77, 8 (2022), 894.

[4] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* 20, 1 (2019), 1–68.

[5] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. 2011. Context in emotion perception. *Current directions in psychological science* 20, 5 (2011), 286–290.

[6] Nicole Betz, Katie Hoemann, and Lisa Feldman Barrett. 2019. Words are a context for mental inference. *Emotion* 19, 8 (2019), 1463.

[7] Jeffrey A Brooks and Jonathan B Freeman. 2018. Conceptual knowledge predicts the representational structure of facial emotion perception. *Nature human behaviour* 2, 8 (2018), 581–591.

[8] Tiffany Doan, Desmond C Ong, and Yang Wu. 2024. Emotion understanding as third-person appraisals: Integrating appraisal theories with developmental theories of emotion. *Psychological Review* (2024).

[9] Bernd Dudzik, Tiffany Matej Hrkalovic, Chenxu Hao, Chirag Raman, and Masha Tsfasman. 2024. Indeterminacy in Affective Computing: Considering Meaning and Context in Data Collection Practices. In *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. 181–185. doi:10.1109/ACIIW63320.2024.00036

[10] Bernd Dudzik, Michel-Pierre Jansen, Franziska Burger, Frank Kaptein, Joost Broekens, Dirk KJ Heylen, Hayley Hung, Mark A Neerincx, and Khiet P Truong. 2019. Context in human emotion perception for automatic affect detection: A survey of audiovisual databases. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 206–212.

[11] Paul Ekman. 2014. Expression and the nature of emotion. *Approaches to emotion* (2014), 319–343.

[12] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[13] Yasaman Etesam. 2025. Vision language models for environmental and emotional awareness. (2025).

[14] Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. 2024. Contextual emotion recognition using large vision language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4769–4776.

[15] Maria Gendron, Debi Roberson, Jacoba Marietta van der Vyver, and Lisa Feldman Barrett. 2014. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion* 14, 2 (2014), 251.

[16] Anirudh Goyal and Yoshua Bengio. 2022. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A* 478, 2266 (2022), 20210068.

[17] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 1 (2012), 723–773.

[18] Matthew Groh and Rosalind Picard. 2021. Context in automated affect recognition. In *Proceedings of the 35th Conference on Neural Information Processing Systems*. 6–14.

[19] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. 2024. Machine learning with a reject option: A survey. *Machine Learning* 113, 5 (2024), 3073–3110.

[20] Sean Dae Houlihan. 2022. *A computational framework for emotion understanding*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[21] Sean Dae Houlihan, Max Kleiman-Weiner, Luke B Hewitt, Joshua B Tenenbaum, and Rebecca Saxe. 2023. Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A* 381, 2251 (2023), 20220047.

[22] Sean Dae Houlihan, Desmond Ong, Maddie Cusimano, and Rebecca Saxe. 2022. Reasoning about the antecedents of emotions: Bayesian causal inference over an intuitive theory of mind. In *Proceedings of the annual meeting of the cognitive science society*, Vol. 44.

[23] Daniel Kahneman. 2011. Thinking, fast and slow. *Farrar, Straus and Giroux* (2011).

[24] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2019. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence* 42, 11 (2019), 2755–2766.

[25] Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. 2023. Additive decoders for latent variables identification and cartesian-product extrapolation. *Advances in Neural Information Processing Systems* 36 (2023), 25112–25150.

[26] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10143–10152.

[27] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. 2022. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8046–8056.

[28] Divyat Mahajan, Mohammad Pezeshki, Charles Arnal, Ioannis Mitliagkas, Kartik Ahuja, and Pascal Vincent. 2024. Compositional risk minimization. *arXiv preprint arXiv:2410.06303* (2024).

[29] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14234–14243.

[30] Desmond C. Ong. 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)*. IEEE, 1–8. doi:10.1109/acii52823.2021.9597441

[31] Desmond C Ong, Jamil Zaki, and Noah D Goodman. 2015. Affective cognition: Exploring lay theories of emotion. *Cognition* 143 (2015), 141–162.

[32] Desmond C Ong, Jamil Zaki, and Noah D Goodman. 2019. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science* 11, 2 (2019), 338–357.

[33] Suzanne Oosterwijk, Kristen A Lindquist, Eric Anderson, Rebecca Dautoff, Yoshiya Moriguchi, and Lisa Feldman Barrett. 2012. States of mind: Emotions, body feelings, and thoughts share distributed neural networks. *NeuroImage* 62, 3 (2012), 2110–2128.

[34] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[35] Ira J Roseman and Craig A Smith. 2001. Appraisal theory. *Appraisal processes in emotion: Theory, methods, research* (2001), 3–19.

[36] Philipp V Rouast, Marc TP Adam, and Raymond Chiong. 2019. Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing* 12, 2 (2019), 524–543.

[37] James A Russell. 1997. into faces: Resurrecting a dimensional-contextual perspective. *The psychology of facial expression* (1997), 295.

[38] Stuart Russell. 2022. Human-Compatible Artificial Intelligence.

[39] Klaus R Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.

[40] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proc. IEEE* 109, 5 (2021), 612–634.

[41] Harald T Schupp, Jessica Stockburger, Maurizio Codispoti, Markus Junghöfer, Almut I Weike, and Alfons O Hamm. 2007. Selective visual attention to emotion. *Journal of neuroscience* 27, 5 (2007), 1082–1089.

[42] Amy E Skerry and Rebecca Saxe. 2014. A common neural code for perceived and inferred emotion. *Journal of Neuroscience* 34, 48 (2014), 15997–16008.

[43] Amy E Skerry and Rebecca Saxe. 2015. Neural representations of emotion are organized around abstract event features. *Current biology* 25, 15 (2015), 1945–1954.

[44] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems* 26 (2013).

[45] Chai M Tyng, Hafeez U Amin, Mohamad NM Saad, and Aamir S Malik. 2017. The influences of emotion on learning and memory. *Frontiers in psychology* 8 (2017), 235933.

[46] Cédric Villani et al. 2008. *Optimal transport: old and new*. Vol. 338. Springer.

[47] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. 2019. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–37.

[48] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion* 83 (2022), 19–52.

[49] Sofia Wenzler, Sarah Levine, Rolf van Dick, Viola Oertel-Knöchel, and Hillel Aviezer. 2016. Beyond pleasure and pain: Facial expression ambiguity in adults and children during intense situations. *Emotion* 16, 6 (2016), 807.

[50] Jiayi Eurus Zhang, Joost Broekens, and Jussi Jokinen. 2024. Modeling Cognitive-Affective Processes with Appraisal and Reinforcement Learning. *IEEE Transactions on Affective Computing* (2024).

[51] Jiayi Eurus Zhang, Bernhard Hilpert, Joost Broekens, and Jussi PP Jokinen. 2024. Simulating Emotions With an Integrated Computational Model of Appraisal and Reinforcement Learning. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–12.