Toward Explainable Affective Computing: A Review

Karina Cortiñas-Lorenzo and Gerard Lacey

***Abstract*— Affective computing has an unprecedented potential to change the way humans interact with technology. While the last decades have witnessed vast progress in the field, multimodal affective computing systems are generally black box by design. As affective systems start to be deployed in real-world scenarios, such as education or healthcare, a shift of focus toward improved transparency and interpretability is needed. In this context, how do we explain the output of affective computing models? and how to do so without limiting predictive performance? In this article, we review affective computing work from an explainable AI (XAI) perspective, collecting and synthesizing relevant papers into three major XAI approaches: premodel (applied before training), in-model (applied during training), and postmodel (applied after training). We present and discuss the most fundamental challenges in the field, namely, how to relate explanations back to multimodal and time-dependent data, how to integrate context and inductive biases into explanations using mechanisms such as attention, generative modeling, or graph-based methods, and how to capture intramodal and cross-modal interactions in post hoc explanations. While explainable affective computing is still nascent, existing methods are promising, contributing not only toward improved transparency but, in many cases, surpassing state-of-the-art results. Based on these findings, we explore directions for future research and discuss the importance of data-driven XAI and explanation goals, and *explainee* needs definition, as well as *causability* or the extent to which a given method leads to human understanding.**

***Index Terms*— Affective computing, explainable AI (XAI), multimodal machine learning, review.**

I. Introduction

SINCE Ancient Greece, it has been widely acknowledged that humans seek explanations in an attempt to understand the world [1]. This ubiquitous search for answers and explanations is inherent to human nature and fundamental to integrate technology into everyday lives. In this context, new personalized and human-driven approaches are driving a paradigm shift from the Internet of Things to the Internet of People, where the focus is not on devices or infrastructure but rather on people [2], [3]. In this new paradigm, human-centric technology must be social and enrich the user experience by being context-aware, allowing for empathetic interactions that are personalized to a given user in a given situation at a given point in time. Furthermore, for technology to be successfully leveraged, interactions not only need to be context-dependent but also predictable [4]. Hence, to enable such scenarios, two key requirements must be met: 1) technology must be able to recognize, understand and express human emotions and 2) in order to be predictable, technology must be explainable and able to expose to the user the reasoning behind its operations in a human-understandable way.

While research on affective computing has witnessed tremendous progress in the past decade [5], [6], research on interpretable AI methods for affect recognition is still in its infancy. Multimodal learning approaches leveraged in the field are usually a black box, offering very little transparency about the reasoning behind their predictions. With most state-of-the-art (SOTA) methods being reliant on deep learning (DL) approaches, opening this black box in affective computing poses some unique challenges.

1) *Multimodal Analysis:* Our experience of the world, including both the perception and the synthesis or arousal of emotions, is inherently multimodal [7], [8]. This means that interpretable methods for affective computing must be able to estimate the relative importance of not only unimodal inputs but also different multimodal sources and their interactions [9].
2) *Real-Time Analysis:* Human emotions unfold over time [10], [11], [12]. We interact with others in real time, and our emotions vary dynamically with temporal context. As a crucial component of real-world affect, time must be taken into account in the interpretation of emotion recognition. Hence, explainable methods must be able to disentangle the importance of different modalities and inputs at different points in time.
3) *Context:* Emotions are influenced not only by temporal context but also by interaction dynamics, the semantics of verbal utterances, environmental context, and social and cultural context. These variables affect the interpretation of nonverbal behaviors and are a fundamental source of interpersonal and intrapersonal variance [13], [14], [15]. Because different behaviors can have different interpretations depending on the context, contextual information specific to the given task should be included when possible in the interpretation of affect recognition.
4) *Ground Truth:* Since affective states are internal to a given individual, we can only observe symptoms of

Manuscript received 10 May 2022; revised 4 December 2022 and 6 March 2023; accepted 19 April 2023. *(Corresponding author: Karina Cortiñas-Lorenzo.)*

Karina Cortiñas-Lorenzo is with the School of Computer Science and Statistics, Trinity College Dublin, Dublin, D02 W272 Ireland (e-mail: cortinak@tcd.ie).

Gerard Lacey is with the Department of Electronic Engineering, Maynooth University, Kildare, W23 A3HY Ireland (e-mail: Gerry.Lacey@mu.ie).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3270027.

Digital Object Identifier 10.1109/TNNLS.2023.3270027

the inferred emotional state or apparent emotions [16]. As a consequence, ground truth is difficult to establish in affective computing [17], and labeling of apparent emotional states can be prone to bias [18], [19], [20]. In this context, explainable methods can help expose not only the model reasoning but also uncover potential biases in the training set.

5) *Ethics:* Emotional states are private; therefore, affective computing naturally raises ethical concerns. In this field, explainable methods are fundamental not only to provide transparency to the user but also to ensure that emotion recognition systems are compliant with privacy laws and can be appropriately audited [21], [22], [23].

Despite these challenges, affective computing has been deployed in diverse domains, ranging from intelligent tutoring systems [24] to health monitoring tools [25]. This diversity of domains and application contexts entails different tasks, users, and risks. For example, failing to recognize a given affective state can have catastrophic consequences when the outcome affects the health of an individual (e.g., mental health diagnosis and monitoring systems [26]) or when a wrong prediction prompts the system to deliver inappropriate interventions (for instance, inappropriate pedagogical interventions in tutoring systems can have a long-lasting impact on people's learning, development, and life-long functioning [27]). On the other hand, when deployed to enhance human–machine interactions, recognition errors can negatively impact user experience, while having negligible harm.

In domains such as education or healthcare, where failing to generalize can deeply impact human lives, explainability is critical [28]. In this type of scenario, explanations provide multiple benefits. First, explanations can help users better understand the system's behavior, nurturing user's trust in decisions when advice or feedback is given. For example, in tutoring and learning systems, understanding feedback is essential to achieve pedagogical success and fundamental for sustained adoption [27]. Second, explanations can provide insights into multimodal learning, allowing researchers and developers to better understand when the model fails to generalize and the key areas for improvement [29]. When deployed on biased training sets, interpretability and model transparency can also help to identify biased labeling patterns in the input data, allowing us to trace back key attributes influencing labeling trends in the training set. In scenarios such as automatic recruitment, this aspect is fundamental to the successful deployment of affective decision support systems [30], [31], [32], [33]. Finally, explanations enable post hoc understanding, facilitating audit analyses to ensure compliance with ethical and privacy regulations [34].

Several surveys exist covering recent advances in affective computing [5], [6], explainable AI (XAI) [35], [36], [37], [38], [39], and multimodal learning [29], [40]. However, to the best of our knowledge, this is the first attempt to review work on XAI with a focus on affective computing. Our goal is to study whether explaining the output of affective computing systems is possible while maintaining SOTA predictive performance, helping researchers in the field grasp the most important aspects of different methods and highlighting the fundamental challenges to be addressed in future research. To this end, we have reviewed over 200 publications on the application of explainable methods in affective computing tasks from high-impact computer science conferences and journals across the fields of affective computing, visualization, and machine learning. With a focus on those applications where SOTA results were achieved or even improved, a smaller subset of these works has been selected to be presented in this article.

The key contributions of this article are given as follows.

1) We collect and synthesize work on explainable affective computing modeling and other related human-centric tasks, leveraging a taxonomy to help researchers efficiently comprehend the strengths and weaknesses of different methods.
2) We provide evidence supporting the hypothesis that explaining the output of affective computing systems does not necessarily lead to worse predictive performance.
3) We discuss the key challenges to be addressed in affective computing and the key requirements differentiating XAI needs in the field from other domains.
4) Based on this literature analysis, we identify and discuss future research opportunities and key gaps to be addressed in order to improve the transparency of affective computing systems in real-world applications.

This article is structured as follows. We start by defining terminology, presenting evaluation methods and the taxonomy to be used in this review (see Section II). Next, we discuss recent XAI developments in the field and group them into three key categories: premodel (see Section III), in-model (see Section IV), and postmodel (see Section V). In Section III, we explore three challenges in premodel understanding: exploratory data analysis (EDA) in multimodal settings, standardization of dataset descriptions and quality tests, and privacy. In Section IV, we review attention mechanisms (see Table I), generative modeling techniques (see Table II), and graph-based methods (see Table III). In Section V, we review postmodel XAI techniques applied on discrete sets of inputs and handcrafted features (see Table IV) and discuss their main advantages and weaknesses. Finally, key challenges and directions for future research are discussed in Section VI, and final conclusions are drawn in Section VII.

II. Problem Definition and Terminology

A. Problem Statement and Evaluation Methods

Can we explain the output of multimodal affective computing systems while preserving model performance? This article explores this question empirically by reviewing past attempts in the literature. Motivated by the existing tradeoff between predictive performance and explainability in XAI and by the specific challenges of multimodal affective computing (see Section I), we break our central research question into the following pieces.

1) *Data Understanding:* The goal of multimodal DL approaches is extracting and learning patterns from training corpora. Hence, understanding the data the system

is working with is crucial to assess the reliability of predictions and improve task performance. Furthermore, enabling human experts to interactively explore the data for informed decision-making can facilitate the integration of human feedback into the modeling loop [41]. From this perspective, we address the following question: how can we explain the output of affective systems in light of the data being used?

2) *Model Understanding (a.k.a. Global Explainability):* Given some inputs and explanatory features, where do model outputs come from? Model understanding deals with grasping how the data get processed during training, explaining how the model behaves on average for a given dataset. More specifically, we want to identify overall feature importance trends and quantify the importance of features and their interactions over the whole affective dataset [9].
3) *Predictions Understanding (a.k.a. Local Explainability):* Given an individual sample, where does a specific prediction come from? Instead of understanding averaged or global trends, in this case, we want to quantify the impact of explanatory features and/or inputs on a given model prediction. The goal is to give the user the ability to question the recommendations from an affective computing system, exposing key features influencing one specific outcome.

B. Terminology and Scope: What Is an Explanation?

Because explanations necessarily involve a human factor, the concept has not yet been formalized in the literature, and there is still debate over what constitutes an explanation and what makes an explanation effective and sufficient for a given user or explainee [36]. In this article, we understand explanations as a form of social interaction, where someone, or some system, provides relevant information to some explainee(s) or human(s) [42]. In this context, the relevancy of the information provided is linked to the knowledge being transferred in the interaction and the ability of the explanation to enhance the explainee's understanding. Hence, in the AI domain, this social interaction involves questions formulated by the explainee and answers or explanations generated by the system. While different types of questions can be answered by explanations, in this article, the main focus is on understanding the *what*: what inputs are related to model outputs and what specific features, interactions, and/or input information lead to and influence a given model prediction. Thus, our scope is the first ladder of explanatory reasoning [43]. We want to understand whether achieving this first stage of explanatory behavior is even possible today without negatively influencing model performance. Other reasoning such as interventionist (*how-questions*) or counterfactual explanations (*why-questions*) are out of the scope of this review.

When it comes to defining the terms interpretability and explainability, some works in the literature refer to interpretable models as models whose reasoning processes are understandable to humans, either because they are inherently interpretable [44] or because certain model properties and components can be directly assigned an interpretation [45], and as explainability as those techniques aimed at explaining black-box models [44], [46]. Others, however, adopt different distinctions between the terms [38], do not differentiate between both concepts [34], [37], [39], or use other terms such as "intelligibility" and "understandability" interchangeably [47]. In this article, we follow the formulation proposed in [41] and understand interpretability and explainability as identical concepts relating to the science that deals with the technical implementation of transparency and traceability in AI approaches, comprehending what a model has done and which part(s) of the model structure influence its recommendations. Because XAI explanations obtained through these approaches do not necessarily lead to user understanding, we use a second concept to capture the human dimension of explanations: *causability*. A notion introduced and discussed in [48], *causability* is about measuring the usability of a given explanation and evaluating whether the information provided meets understanding goals in a given context, helping a human connect the recommendations to a mental model and giving the user the power to understand, confirm, or overrule model outputs. Because *causability* refers to a human model, it necessarily involves the evaluation of the quality of explanations.

Inspired by the ideas proposed in [41], in addition to assessing the ability of the methods to enable data understanding, model understanding, and/or predictions understanding, the work reviewed in this article is also evaluated against two additional dimensions: first, the ability of the proposed technique to provide a *technical explanation* while maintaining or improving SOTA results on a given affective task, and second, the evaluation of the *causability* of the generated explanation and the extent to which the proposed method leads to human understanding.

C. Affective Computing and Similar Fields

Since first introduced in the late 1990s, affective computing research has evolved over the years from having a single focus on human effect or emotions [16] to modeling social signals and nonverbal behavior in human multimodal interactions [49], [50], as well as modeling other phenomena, such as personality traits [51], cognitive abilities [52], [53], or psychological disorders [26]. Despite taking many forms, all of these variants involve the same set of fundamental characteristics outlined in Section I: the use of multimodal time-dependent data, the sensitivity to context, the difficulty in establishing ground truth, and several ethical concerns involving the use of these technologies in different real-world scenarios. Hence, in this review, we not only consider work focused on predicting emotion labels but also include other human-centric tasks, such as modeling psychological and neuro disorders, mental states, personality traits, cognitive abilities, social signals, and sentiment analysis.

D. Taxonomy

In this article, inspired by the taxonomy proposed in [34], we group different XAI methods for multimodal affective computing considering *when* they are applied.

1) *Premodel (Data Understanding):* The goal of premodel techniques is to understand the data used by the system before the actual training takes place. Relating explanations back to input data is a fundamental step toward understanding the underlying model mechanisms, helping a human connect model shortcomings with the limitations of the training datasets.
2) *In-Model (Model Understanding):* In-model XAI techniques are focused on providing transparency into how the data get processed during training. By understanding how the model learns from data and how relationships between modalities and time influence the final outcome, the goal is to trace back predictions to inputs and learned features.
3) *Postmodel (Predictions Understanding):* Postmodel XAI techniques are applied after training and either consider the model as a black box, using *explainer* mechanisms to relate inputs to outputs, or leverage some aspect of the model's internal structure. Traditionally applied to unimodal data in the XAI domain, in this article, we review attempts to incorporate these techniques into multimodal affective settings while highlighting their key advantages and limitations.

III. Premodel XAI Methods for Affective Computing

Data-driven systems can generally improve their predictive performance by adjusting their training algorithms and/or improving the quality of the training dataset. In affective computing, the latter requires both an exploration of multimodal datasets and a common framework against which to measure and quantify data quality. Understanding the data used to train affective algorithms is not only beneficial to improve modeling performance but also necessary to enable model transparency. If we consider explanations as social interactions, generating explanations should be framed as a conversation, where a user can query the system in an interactive way as new answers are provided [42], [54]. Premodel XAI methods can not only facilitate this task by enabling interactive data exploration for enhanced decision-making but can also help the system relate explanations back to data. For instance, when constrained and unrepresentative training datasets limit the ability of the model to generalize, unexpected predictions on unseen cases can be explained on the basis of the input dataset quality. In addition, because data-driven affective systems are fundamentally exploiting associative patterns in the data, relating explanations back to the input data can help researchers better understand and identify potential limitations of the system and areas of improvement from a data-centric perspective.[1]

Applied to data before training, the methods reviewed in this section are aimed at enabling a common understanding of multimodal affective data. Rather than transforming a given dataset to improve its quality, the goal of the methods discussed is to gain insights into the data used for learning and/or inference in order to explain model outputs. Our discussion is centered on three main challenges.

[1]https://datacentricai.org/

1) *EDA for Multimodal and High-Dimensional Data:* Exploring multimodal datasets for enhanced data understanding involves dealing with high-dimensional data. Techniques to reduce both the time and space complexity of the EDA task and methods to project high-dimensional data into lower dimensional representations are needed.
2) *Lack of Standardized Affective Dataset Descriptions and Quality Tests:* A common understanding of datasets in the field and tests to assess their limitations can accelerate multimodal affective research and deployment into real-world applications. However, creating, documenting, maintaining, and evaluating multimodal datasets are expensive, and disjoint efforts on these fronts make standardization a challenge.
3) *Privacy:* Due to the nature of affective computing data, premodel explainability can unintentionally lead to a leak in personal and sensitive data, potentially infringing on human rights. In addition, in real-world scenarios, both training sets and data models may be proprietary, limiting the availability of assets for explainability and increasing the risks of intellectual property infringement.

A. EDA: Understanding Multimodal Affective Datasets

The goal of EDA is to discover patterns in the data to foster hypothesis development and refinement. Deep-rooted in statistics, EDA has an emphasis on understanding the data through visual exploration and metrics. While being a well-established statistical tradition, EDA can become challenging in fields such as affective computing, where dataset dimensions are usually high and different modalities can be noisy, missing, or corrupted. The high dimensionality of data in this domain requires projections into lower dimensional, more interpretable, representations while preserving the underlying structure. In this context, building lower dimensional data representations that encapsulate meaning is a key challenge. Moreover, since emotions are time-dependent, volumes of data can grow even further, increasing computational complexity. This dependence on time, involving the use of multimodal sequences of unstructured data, also entails significant difficulties when aligning different modalities for visual comparison.

Due to these challenges, past work on affective data visualization has mainly focused on understanding affective annotations rather than input data. For instance, Zeng et al. [55] built EmoCo, an interactive visual analytics system to analyze emotion coherence and target alignment across different modalities in public speaking videos, Zeng et al. [56] introduced EmotionCues, an exploratory tool to visualize and explore the affective state of students in a classroom, and Wang et al. [57] presented DeHumor, a visual analytics application for analyzing humor behaviors in public speaking. Understanding the target distribution can also help improve the reliability of annotations. Wang et al. [58] explore the distribution of emotion annotations using outlier detection methods and use these insights to correct outliers toward the learned distribution, reducing labeling noise and outperforming previous SOTA results. Escalante et al. [59] analyze different aspects

of the target variable in the First Impression dataset, including intravideo and intervideo variance, and use these insights, together with studies of the correlations with sensitive traits, such as gender and ethnicity, to uncover existing biases in the dataset.

Analyzing patterns in the input data can also be beneficial to understand sources of interpersonal and intrapersonal variance. For example, Carbonell et al. [60] use unsupervised methods to augment the supervised analysis of emotion expressions, revealing additional insights into how emotion expressions differ by individual and gender. Also aimed at comparing and contrasting data across emotions and individuals, the authors of AffectiveTDA [61] propose to use topological data analysis (TDA) to build an explainable visual data representation of facial expressions using facial landmarks.

Unsupervised methods such as clustering can not only help improve predictive performance by identifying the most relevant discriminant features [62], [63] but can also ease feature selection for semantic interpretation. For instance, in [64], singular value decomposition-based coclustering is used to identify the most salient regions in face images. Other unsupervised techniques, such as PCA, can also facilitate dataset understanding and comparison: in [65], a corpus similarity measure based on PCA-ranked features is proposed to identify similar emotion datasets, helping practitioners select and merge datasets for speech emotion recognition. Although the high dimensionality of multimodal data in affective computing can hinder the implementation of traditional EDA techniques, the computation of simple metrics, such as similarity metrics, can still provide explainability benefits while remaining computationally feasible. For instance, in [66], dynamic time warping (DTW) is used to compute in real time the similarity between two multimodal temporal sequences. Hierarchical clustering is then applied to these similarity values to identify abnormal nonverbal behaviors in the participants of a TV show. Identifying outliers and measuring the dissimilarity between observations using techniques such as these ones could also be useful to explain abnormal model behavior (e.g., unexpected predictions due to distribution shifts).

B. Standardization of Affective Dataset Descriptions

Affective computing research has suffered from a lack of available datasets due to expensive labeling, requiring multiple annotators to minimize annotation bias [17]. However, as more datasets are created, the lack of standardized dataset descriptions and quality tests can hinder the field to scale into real-world applications. In this context, descriptions and tests can help practitioners understand and assess the quality of the training data, providing them with tools to analyze the tradeoffs of different datasets while allowing them to explain model behavior in light of the input data. In the discussion that follows, we understand **dataset descriptions** as metrics that provide transparency into what a given dataset contains, including aspects such as the distribution statistics for input features, the assessment of the labels' construct validity, or the justification of the annotation scheme. On the other hand, **quality tests** provide an assessment of the external consistency of the data against a given quality framework (e.g., bias tests [67]). Whereas dataset descriptions help answer the question *what does the dataset contain?*, quality tests address a value-based one: *is this dataset good enough for this task?*

1) Dataset Descriptions: In general, when describing an affective computing dataset, different metrics are usually reported in the literature. While the number of individuals and the distribution of the target labels are commonly informed, different authors consider incorporating additional distinct metrics. For instance, the authors of SEWA [68] indicated the age range, gender, and culture of participants, while, in CMU-MOSEI [69], only the gender is informed. Most datasets contain English recordings and ignore accents, but, in RECOLA [70], the authors presented a multimodal corpus of interactions in French and reported the distribution of mother tongues of the participants. Annotation schemes and ways to ensure construct validity are also diverse: using the same examples, to build SEWA, five annotators were hired for each cultural background, whereas the authors of RECOLA relied on a total of six annotators and used normalization techniques to reduce variability in human judgment. While annotators of RECOLA could annotate video sequences from another dataset to become familiar with the annotation interface, the authors of CMU-MOSEI provided a 5-min training video to crowdsourced judges from Amazon Mechanical Turk. Although some psychological constructs require a minimum time to assess and cannot have continuous metrics (e.g., personality traits), annotation schemes also differ even for the same psychological constructs. For example, SEWA and RECOLA use continuous measures of valence and arousal, while other datasets use different discrete labels for a single video, frame, or utterance (e.g., CMU-MOSEI).

Standardization of these aspects is a challenge in the field, making it hard for researchers and practitioners to effectively choose a multimodal dataset and incorporate some of these characteristics into explanations. Databases such as the ones mentioned above are usually manually collected to fulfill the requirements of specific research tasks, resulting in disjoint efforts when reporting dataset descriptions. To the best of our knowledge, in affective computing, there are no standards on how to document datasets yet. In the broader machine learning field, however, several approaches already exist. For instance, Datasheets for Datasets [71] propose to document not only the dataset composition but also its purpose, collection process, and recommended uses, including questions appropriate for both academic researchers and product teams. In production systems, Data Cards [72] are an alternative framework aimed at capturing critical information about a dataset across its life cycle. Similar approaches also exist for structured data [73] and domain-specific fields, such as natural language processing (NLP) [74]. Some of these existing dataset documentation frameworks could be leveraged in affective computing, mitigating potential ground-truth biases while preventing researchers and practitioners from using datasets in the wrong contexts.

2) Quality Tests: Assessing the quality of a given dataset is also part of understanding the data, as it helps identify its limitations and strengths for a given task. In this regard,

due to the difficulty of establishing ground truth, bias is an important quality aspect in affective data. Booth et al. [67] provide guidelines to approach this topic in the field, including a framework to identify sources of bias when inferring psychological constructs and metrics to quantify fairness at different stages of the machine learning pipeline.

Other quality dimensions, such as the completeness of different modalities and noise topologies, also lack standardized tests in affective computing. In the broader machine learning community, however, Datasheets for Datasets [71] encourage dataset creators to inform about any errors, sources of noise, or redundancies in the data. In addition, in the context of real-world deployments where data can be unpredictable and ever-changing, quality assessments for new data are crucial to understand and explain the behavior of retrained models. Despite its ubiquity in real-world scenarios, identifying and measuring data drift in multimodal affective computing are challenging: both changes in input data and target attributes can be unknown [75], drift can happen at different levels [76], and it can involve both unimodal and cross-modal joint distribution shifts [77].

Dataset quality is also dependent on the task: the quality of a given dataset can be adequate for a given task but poor in a different context. Documenting the purpose of a given dataset, as well as the recommended and not recommended uses [71], can help practitioners and researchers better select datasets, increasing their awareness of dataset limitations, and ultimately helping them understand model behavior in light of the data being used. As affective computing research evolves, standardization of dataset documentation and quality tests is a step toward scaling the deployment of affective computing systems in real-world applications.

C. Privacy

A key challenge in responsible AI is the tension between explainability and privacy goals [78], [79]. In research settings, how a dataset is collected, including its primary source of information and the level of consent required from participants, determines the amount of information that can be shared. In real-world production systems, the training data may not be available due to privacy and/or security issues, and both models and datasets may also be proprietary. In these contexts, revealing information through explanations could infringe both privacy and intellectual property rights [80], [81].

Since affective states are latent, it can be safer to assume that users of real-world affective computing systems will consider their data private by default [23]. By adopting the default position that affective data are only for personal consumption, federated learning (FL) can help protect privacy while allowing for local premodel explainability. Following this approach, all the training is done on-device, and only model parameters are sent to a centralized server [82], [83]. In this setting, premodel explainability could leverage only local data on-device, helping a user understand how their own data influences the overall system behavior. Although, in affective computing, FL has already been used (e.g., stress-level monitoring [84], pain estimation [85], and mood prediction [86]), explainability in these types of scenarios remains largely unexplored. One key challenge relates to explaining the system behavior locally when this depends on both the inference logic of local models and the influence of multiple local models impacting the overall master model(s). While premodel explainability could help a user understand the influence of their own data and patterns locally, understanding the overall system behavior opens a broader challenge in the explainability space, namely, how to extract explanations from a complex system in which multiple models and blocks interact with each other.

When data need to be transferred to a centralized server, how the data are collected and used, as well as any risks and limitations, should be communicated to users transparently [23], [87]. In this type of scenario, differential privacy and data obfuscation strategies could be useful to preserve users' privacy while maintaining the relevant predictive patterns in the data. In affective computing, however, disentangling personal data from sensitive attributes can be extremely hard. To overcome this challenge, recent techniques in video-to-video synthesis could be leveraged. For example, Haddonll et al. [88] explore the idea of transferring visual behaviors onto a target video of nonexistent anonymous faces displaying the same apparent emotion, offering a baseline for the suitability of expression transfer and anonymization in affect prediction models. In speech emotion recognition, Feng et al. [89] propose a privacy-preserving data transformation technique to remove sensitive information in the training data, preventing inference of sensitive attributes while preserving data utility, and in [90], different approaches to obfuscate user identity in multimodal mobile data are presented. While these pieces of work address how differential privacy and data obfuscation approaches could be implemented, none of them explore the explainability of the final anonymized datasets. The methods discussed in Sections III-A and III-B could be helpful not only to allow model developers and researchers better understand model behavior in light of the used data but also to assess and evaluate premodel explainability before and after anonymization tasks.

D. Premodel Methods' Conclusion

Traditionally focused on understanding a given modeling algorithm, XAI methods have largely ignored data understanding. However, as discussed in this section, a big part of understanding the behavior of a data-driven system involves understanding the data that it is based on. Premodel explainability can help researchers and model developers understand multimodal relationships in the data, the influence of context, and sources of variance, and evaluate the reliability of target annotations in affective data. In real-world applications, premodel explainability can help end users understand the outputs of an affective computing system and assess their relevancy in a particular context. Specifically, when the explanations are linked to a given user-generated data, this understanding can provide agency to users, giving them the ability to control the system behavior by adjusting the data being fed.

However, several challenges lie ahead. First and foremost, understanding multimodal and time-dependent data is not

TABLE I
SUMMARY OF ATTENTION-BASED XAI METHODS FOR AFFECTIVE COMPUTING TASKS

In-Model XAI Methods for Affective Computing - Attention Mechanisms		
Task	**Reference**	**Multimodal**
Sentiment Analysis	Tsai et al., 2019 [95]	✓
Emotion Recognition	Wang et al., 2019 [96], Tzirakis et al., 2021 [97], Gu et al., 2018 [98] Nguyen et al., 2021 [103]	✓ ✗
Mental health	Ahmed et al., 2022 [100], Han et al., 2022 [101]	✗
Conversation Analysis	Gu et al., 2018 [99]	✓
Automatic Hiring	Hemamou et al, 2021 [102]	✓

TABLE II
SUMMARY OF GENERATIVE-BASED XAI METHODS FOR AFFECTIVE COMPUTING TASKS

In-Model XAI Methods for Affective Computing - Generative Modelling			
Methodology	**Task**	**Reference**	**Multimodal**
Probabilistic Graphical Models			
HMMs	Sentiment Analysis	Perikos et al., 2021 [116]	✗
Bayesian Networks	Engagement prediction	Heimerl et al., 2020 [15]	✓
Deep Generative Models			
Deep Markov Model	Not applied yet in affective computing	Zhi-Xuan et al., 2020 [122]	✓
Multimodal Factorization	Personality Trait Recognition, Sentiment Analysis, Emotion Recognition	Tsai et al., 2018 [123]	✓
Probabilistic Programming			
Hybrid methods	Emotion Recognition	Ong et al., 2019 [124]	✗

straightforward and, in some situations, may not even be computationally feasible. Nonetheless, depending on the application, full comprehension of the datasets may not be necessary. Simple metrics, such as similarity metrics, could still be valuable and worth including in explanations. In addition, standardization of dataset metrics and documentation, potentially adapting frameworks, such as Datasheets for Datasets [71], could help practitioners and researchers better understand the behavior of their developed systems in light of the data being used. Last but not least, premodel explainability needs to be aligned with privacy goals. Potential avenues to preserve privacy while incorporating premodel explainability include FL approaches, as well as differential privacy and data obfuscation methods. Since these techniques add additional complexity, further research is needed to move from a single-algorithm/modeling-centric approach to XAI to a holistic approach that considers not only the data being leveraged by the system but also the different components and blocks in it.

IV. IN-MODEL XAI METHODS FOR AFFECTIVE COMPUTING

Often referred to as interpretable ML [44], [45], the goal of in-model XAI methods is to provide transparency into how input data get processed and transformed into model predictions. Rather than incorporating additional mechanisms post hoc to understand the relationship between model inputs and outputs, in-model techniques are inherent to model architecture and take an active part in model training, contributing toward both model understanding and prediction understanding. In affective computing, these transparent modeling mechanisms aim to answer the following questions.

1) What is the impact of different behavioral cues and multimodal features in model predictions?
2) How important are cross-modal interactions and intramodal features in predicting a given target?
3) How does time influence the relationship between input features and model outputs?

In order to answer the questions above, in-model XAI techniques in affective computing have a focus on how to disentangle feature interactions and how to effectively fuse and align different modalities. In the discussion that follows, we review past work in the field implementing the following methods: attention mechanisms (see Table I), generative modeling (see Table II), and graphs (see Table III).

A. Attention Mechanisms

The aim of attention mechanisms is to enhance or *attend to* the most relevant parts of the input data while ignoring or putting less weight on irrelevant information for a given prediction task. In its simplest form, attention produces a weighted combination of input vectors, where the most relevant inputs to predict a target are assigned higher weights and the least relevant ones smaller weights [91]. Formally, we can think of attention mechanisms as mimicking the retrieval of a value v_i (e.g., an input embedding vector) for a query q (e.g., a contextual embedding) based on a key k_i (usually the same input embeddings vectors) in a database, where the query outcome is a weighted combination of values v_i based on the

TABLE III
SUMMARY OF GRAPH-BASED XAI METHODS FOR AFFECTIVE COMPUTING TASKS

In-Model XAI Methods for Affective Computing - Graphs			
Methodology	**Task**	**Reference**	**Multimodal**
Graph Representations			
Graph Neural Network (GNN)	Emotion correlations	He et al., 2019 [137], Dai et al., 2021 [138]	✗
	Emotion recognition with context modelling	Zhang et al., 2019 [139], Gao et al., 2021 [141], S. Wu et al., 2022 [140]	✗
	ABSA	Zhang et al., 2019 [133], Hou et al., 2021 [134], Pang et al., 2021 [135], R. Li et al., 2022 [136]	✗
	Facial Expression Recognition (FER)	Zhou et al., 2020 [127], Liu et al., 2021 [128]	✗
	Body Emotion Recognition	Ghaleb et al., 2021 [129]	✗
	EEG signal-based emotion recognition	Ye et al., 2022 [131], C. Li et al.2023 [132]	✗
	Emotion Recognition in Conversation	Ghosal et al., 2019 [142], Ishiwatari et al., 2020 [144]	✗
		Lian et al., 2020 [143], Xu et al., 2022 [145]	✓
	Negotiation in Conversation	Joshi et al., 2021 [146]	✗
Dynamic Fusion Graph (DFG)	Sentiment Analysis	Zadeh et al., 2018 [69]	✓
Modal Temporal Attention Graph (MTAG)	Emotion Recognition	Yang et al., 2020 [147]	✓
Knowledge Graphs (KGs)			
COMET ConceptNet	Emotion Recognition in Conversation	Ghosal et al., 2020 [153]	✗
		Wang et al., 2021 [154]	✗
WordNet Manually constructed graph SenticNet, SenticLSTM	ABSA	Zhong et al., 2022 [156]	✗
		Zhao et al., 2021 [155]	✗
		Ma et al., 2018 [162]	✗

similarity of the query vector q and the different key vectors

$$\text{attention}(q, k, v) = \sum_i \text{similarity}(q, k_i) * v_i. \quad (1)$$

Because the weights associated with input embeddings are trainable, these learned parameters can give insight into the specific inputs that the model has considered useful when training. In affective computing, this can be useful to understand the contribution of different modalities and input features at different points in time. Nonetheless, previous research has shown that the relationship between attention weights and model outputs is not necessarily direct [92], [93]. Jain et al. [92] show that attention scores are often uncorrelated with gradient-based feature importance measures, and adversarial attention distributions can be found without altering model predictions. Although attention mechanisms can provide a look into the inner workings of a model by producing an easily understandable weighting of hidden states, attention scores are not unique nor do they necessarily provide a faithful interpretation of the link established by the model between inputs and outputs. Hence, post hoc analyses of attention weights and tests, such as the ones proposed in [94], are needed to make informed decisions about the extent to which model outputs can be interpreted via attention weights.

To the best of our knowledge, in affective computing, there is very little work incorporating post hoc analyses of attention scores. Most work uses attention to improve model performance without forcing an alignment in multimodal data, only visualizing attention activations for qualitative analysis [95], [96], [97], [98], [99], [100], [101]. The authors of Multimodal HireNet [102], however, incorporate a quantitative analysis of attention weights, leveraging attention mechanisms in a hierarchical fashion to understand the importance of questions, moments, and modalities in a hiring interview. The statistical analysis on attention activations includes visualizations of temporal attention for different modalities, unsupervised modeling to extract the regions of maximum attention, and supervised techniques to understand the relevance of attention slices in predicting employability. While they are unable to identify whether the effect of attention slices is positive or negative and *causability* of the generated explanations is not yet addressed, their approach provides insight into not only the salient inputs that the model considers relevant for the prediction task but also the influence of those elements on the final predictive performance.

Because attention scores are the result of a training process on some input data, attention weights give insight into specific information deemed useful when training but are not necessarily correlated with human attention. What the model decides to *attend to* is not necessarily aligned with what a human would *attend to* if given the same inference task. Hence, attention-based explanations can potentially generate misalignment with the user's prior beliefs, leading to untrusted insights that are not relevant to the *mental model* of the explainee [42]. Along these lines, Nguyen et al. [103] explore whether attention weights are aligned with semantic meaning in emotional narrative understanding. While the qualitative analyses imply that attended words carry emotional semantic meaning, the work does not include quantitative analyses nor user evaluations to measure semantic meaning and the extent to which attention activations are aligned with these connotations. In that respect, further research is needed to assess the correlation between model attention and human meaning in different training corpora and when incorporating multimodal data.

B. Generative Modeling

Compared to discriminative models, generative models can generate new data instances by modeling the distribution of

the data itself. Given a set of input data X and a set of targets Y, the goal of a generative model is to approximate the joint distribution of the observed data $P(X, Y)$ by maximizing the likelihood of the data under the model assumptions. Discriminative models, on the other hand, are focused on discriminating label data Y while considering input data X, approximating the conditional probability $P(Y|X)$. Because this latter probability requires less knowledge about the data to be modeled than the joint distribution, discriminative models usually require larger amounts of data to train and perform well. Conversely, to model a joint probability, knowledge about the input data X and label data Y is needed, requiring more assumptions about the underlying structure of the data.

Not all generative models are explainable. For instance, variational autoencoders (VAEs) and generative adversarial networks (GANs) are black box by design. Some generative approaches, however, offer XAI advantages. Generative techniques such as probabilistic graphical models (PGMs) allow the incorporation of expert knowledge via graphs, ensuring alignment between model explanations and explainees' prior knowledge. For example, in healthcare, these models can represent researchers' assumptions about the causal structure among variables [104], and in affective computing, these representations can encode theories of emotion [12], allowing explanations to be interpreted in light of a given theory. In these models, the relationship between nonobserved variables (e.g., internal emotions, beliefs, or desires) and observed variables (e.g., emotional expressions and verbal cues) can be explicitly modeled at different time steps. For example, the evolution of latent emotions, $Z_t \rightarrow Z_{t+1}$, could be modeled using observed emotional expressions X_t and perceived emotions Y_t. This flexible representation, encoding both observed and nonobserved variables in a structure that has intrinsic meaning, allows the interpretation of relationships in the model in light of contextual cues. For example, when modeling engagement [15], contextual variables, such as backchannel or interruptions, and social factors, such as gender, can be explicitly defined as variables in a PGM.

Modeling joint distributions has a long history in affective computing. Before the rise of DL, different works leveraged probabilistic models, such as hidden Markov networks (HMMs) [105], [106], [107], Bayesian networks (BNs) [108], [109], [110], Gaussian mixture Models [111], [112], and Boltzmann machines [113], [114]. Despite providing better explainability, these generative approaches have two key disadvantages.

1) *Representation: Low Capacity to Model Complex Relationships:* Because generative models usually incorporate domain knowledge through a set of inductive biases, these biases are commonly strong simplifying assumptions, limiting their ability to capture real-world complex variability.
2) *Inference: High Computational Complexity:* While inference in generative models is NP-hard, approximate inference methods can generally provide good-enough approximations for many cases encountered in practice [115]. However, when incorporating complexity to increase expressive power, approximate inference methods can grow their computational complexity even further, limiting their ability to train fast.

Several examples of recent work in affective computing have tackled the challenges above. The following discussion is framed around three common approaches in the field: PGMs, deep generative models, and probabilistic programming.

1) PGMs: Joint multivariate distributions over large numbers of random variables can involve the specification of an exploding amount of probabilities [115]. This expensive representation involves high memory complexity, making it cumbersome to manipulate both from a computational perspective and from a human expert standpoint. PGMs reduce representation complexity by exploiting independence properties in the data and leveraging graphs as compact representations of joint probabilities. In affective computing, recent work has leveraged two basic types of PGMs: 1) Markov networks, using undirected graphs with cyclic dependencies and 2) BNs, using a directed acyclical graph and explicitly modeling conditional dependencies.

Starting with the first, Perikos et al. [116] introduce novel HMMs to recognize sentiments in text. HMMs face two key challenges in this type of task: first, the need for a high feature space dimensionality to achieve high capacity, leading to sparsity issues and very small probabilities, and second, the difficulty of generalizing over previously unseen observations. To solve these challenges, Perikos et al. [116] propose a semantic clustering of words and a heuristic approach to perform feature selection via majority voting, as well as a smoothing factor for probability estimates of out-of-vocabulary observations to avoid multiplications by zero. While these new explainable HMM architectures require smaller datasets and lower computational costs, they achieve lower performance compared to SOTA deep neural networks (DNNs). Although no user evaluations are carried out in their work, the authors show how HMMs can qualitatively enable interpretability at every step of their operations, achieving performance levels on the task of sentiment analysis that based on the application could be deemed acceptable.

Compared to Markov networks, BNs' model directed relationships between random variables or nodes in a network. In affective computing, this type of architecture has been used to encode context in a variety of tasks, including engagement prediction [117], affect recognition from speech [118], and emotion modeling in dyadic conversations [119]. To improve their modeling capacity, BNs have also been combined with DL approaches. For instance, Heimerl et al. [15] use DL methods for feature learning from multiple modalities and rely on theoretical knowledge to choose the relevant features to include in the model. While the work in [15] does not include a full comparison of performance to SOTA DL methods, the authors show how these models can offer a direct read on the probabilistic relationships between relevant factors. Nevertheless, further research is needed to evaluate the *causability* of these probabilistic insights.

2) Deep Generative Models: Hybrid approaches combining the benefits of both discriminative and generative models are more and more common in the literature [12]. The general idea

is to leverage DL to learn complex probability functions in generative models, facilitating inference, and parameterization of probability functions while enabling model understanding through uncertainty estimation. For instance, in a deep Markov model (DMM), the probability functions defining the relationships between latent and observable variables are parameterized using neural networks [120], [121]. In affective computing, however, DMMs need to handle multiple data modalities, and therefore, uncertainty-aware multimodal fusion is required. In this regard, Zhi-Xuan et al. [122] introduce a factorized inference method for multimodal DMMs. This multimodal generalization is derived via a factorized variational posterior, and while not tested directly on affective inference tasks, the authors demonstrate that the method is capable of multimodal temporal inference tasks, using both synthetic and real-world multimodal data with varying levels of data deletion.

Also, combining generative and discriminative approaches, Tsai et al. [123] propose to optimize for a joint generative-discriminative objective across multimodal data and labels. The proposed approach factorizes representations into two sets of independent factors: multimodal discriminative factors (shared across all modalities and containing joint multimodal features for discriminative tasks) and modality-specific generative factors (unique for each modality and containing the information required to generate each unimodal data). Because generative factors model the joint distribution for each modality, the incorporation of these into the architecture allows the reconstruction of missing modalities, helping the model achieve SOTA performance on six multimodal datasets. By modeling unimodal variance using a generative approach, the authors are also able to interpret the learned multimodal representations in light of the disentangled unimodal contributions, understanding the influence of each modality toward multimodal predictions at each time step.

3) Probabilistic Programming: Because theory-driven approaches require domain knowledge and need to be hand-tuned to specific tasks and contexts, these methods tend to have larger barriers to adoption. Usually built incrementally and through iterations, being able to quickly compare different theories is key for the sustained use of generative models. Ong et al. [124] propose a probabilistic programming approach to affective computing, modeling psychological-grounded theories as generative models of emotion and implementing them as stochastic, executable computer programs. Because PGMs suffer from limited representation power and computationally intensive inference, probabilistic programming can be seen as a modern successor [125]. Instead of a graphical representation, conceptual knowledge is defined as stochastic programs embedding randomness into their execution. These programs allow for explicit modeling of uncertainty and direct interpretation, as well as modularity and hierarchical setups to model complex phenomena. To jumpstart adoption, Ong et al. introduce different probabilistic libraries leveraging deep-learning for fast inference, such as Pyro, and go through tangible case studies and accompanying code to illustrate the interpretability strengths of these methods [124].

C. Graphs

A major challenge in affective computing involves how to incorporate context into model explanations. While generative models use a theory-driven approach to incorporate context through latent variables, graph representations can integrate context explicitly, encoding both spatial and temporal relationships in the data by using a set of edges or connections between different items or entities. These relationships can be extracted and incorporated into explanations while still leveraging data-driven modeling techniques, overcoming both the interpretability limitation of DNNs and the capacity constraints of generative approaches. Moreover, graphs can not only be used to represent key relationships in unstructured data but also be leveraged to represent *commonsense* reasoning or specific domain knowledge. The latter, commonly referred to as knowledge graphs (KGs) or semantic networks [126], offers key advantages to explainable modeling, including both the incorporation of external knowledge and symbolic logic.

In the discussion that follows, we give an illustrative overview of recent work in affective computing where graphs have been used as a data representation tool to encode task-relevant information. Next, we review past work in the field leveraging KGs and discuss the key advantages and challenges of these methods. Rather than providing an exhaustive collection of references in this area, our goal is to provide the reader with an illustration of what is possible, building knowledge on how graphs can contribute toward explainable affective computing.

1) Modeling Data With Graph Representations: Certain affective computing tasks can benefit from exploiting inherent relational structures in the data. In this type of application, graphs are well suited to improve both predictive performance and in-model interpretability. For example, instead of action units (AUs), muscles in a human face and their relationships could be represented via a graph [127], [128]. By exploiting the dynamics of facial topology, the proposed method in [128] has greater discriminative power and higher interpretability. Along the same lines, Ghaleb et al. [129] present an explainable approach for bodily expressed emotion recognition. Using the body joints of the human skeleton as nodes in a graph, the proposed architecture leverages graph convolutional networks (GCNs) [130] and attention mechanisms to identify which body part contributes the most to emotional inference. Also, using GCNs, in electroencephalogram (EEG) signal-based emotion recognition, Ye et al. [131] propose a hierarchical dynamic approach to learn topological relationships among EEG channels, and Li et al. [132] apply a graph topology feature learning strategy to identify discriminative brain regions and explain which ones relate the most to emotion.

Another task where graphs can exploit inherent structures in the data is sentiment analysis and, more specifically, aspect-based sentiment analysis (ABSA). To identify the sentiment polarity toward different aspects contained in a text, graphs encoding the syntactic tree structure have been used to

disambiguate information [133], [134], [135], [136]. With no specific focus on XAI, the methods proposed in these works rely on graphs to improve model performance by modeling correlations between different syntactic elements in a sentence. To the best of our knowledge, no existing work has attempted to assess the value of these methods for model transparency.

Even when no inherent structures exist in the data, graphs can help overcome other key challenges in affective computing. Starting with the reliability of emotion annotations, graphs have been used to model the correlations between different emotions. Rather than considering a given emotion as a stable single label, the idea is to acknowledge the fuzzy nature of emotions and explicitly model the correlations among different affective states using a graph [137], [138]. As shown in [138], this offers several advantages, including a better understanding of the emotional state of a given individual and a direct interpretation of the modeled graph.

Graph representations can also help modeling context, incorporating contextual cues in the interpretation of affect recognition. For instance, in [139], graphs are used to learn the affective relationships between different context elements in a picture, and in [140], a hierarchical method based on scene graphs is presented. Following a similar reasoning, Gao et al. [141] propose a novel framework based on a GCN that leverages both spatial and temporal contextual features for video emotion recognition, outperforming SOTA methods for context-aware emotion recognition while allowing a direct visualization between final emotion predictions and salient regions in video frames.

Emotion recognition in conversations is another task where contextual factors play an important role. Because modeling emotions in a conversation requires reasoning over long-term dependent contextual information, sequence-based DNNs struggle to capture multiple dependencies over long sequences of time [142]. Hence, graphs have been used to capture both temporal dynamics and intraspeaker and interspeaker dependencies. By modeling the conversation using a directed graph, utterances are represented by nodes and edges capture the dependencies between utterances coming from different speakers. In this way, the entire conversational corpus is symbolized as a heterogeneous graph, and the emotion detection task is framed as a classification problem of the utterance nodes [143]. Ghosal et al. [142] present DialogueGCN and use edge weights to visualize the dependencies between different speakers. Extensions of this work include the incorporation of sequential context [144], an extension of the method to M distinct speakers in a conversation [143], and solutions for long-term contextual information without considering a fixed window size [145].

In cases where conversations have a clear defined goal, graphs can also provide a useful structure to understand conversation dynamics. Joshi et al. [146] propose Dialograph, a text-based negotiation system based on graph attention networks for interpretable prediction of optimal negotiation strategies. By using a graph where nodes correspond to negotiation strategies for a given utterance and directed edges capture the influence between strategies, the proposed architecture allows the interpretation of negotiation strategies via attention. By relying on graphs, Dialograph not only improves the understanding of negotiation strategies in a conversation but also outperforms previous baselines for downstream dialog generation.

Finally, modeling multimodal dynamics using a graph can provide several benefits, including an improved performance by helping the network to explicitly model cross-modal interactions and an improved understanding of the fusion mechanism. For example, Zadeh et al. [69] use edge weights to study how modalities relate to each other across time, and Yang et al. [147] propose modal-temporal attention graph (MTAG), an interpretable graph-based approach that converts unaligned multimodal sequence data into a graph with heterogeneous nodes and edges. Because each node in the proposed graph represents a video, text, or audio utterance, a dynamic pruning method and attention mechanisms are considered to reduce the complexity of interpretation, achieving SOTA performance while allowing for a qualitative interpretation of model outcomes.

2) Knowledge Graphs: KGs are graphs that store factual information or *commonsense* knowledge by using nodes/entities and edges to represent relationships between entities. Typically encoded as a set of rules, knowledge in a KG can be either represented via a graph or by using subject–relationship–object triples [148]. While directly interpretable, these representations make it hard to ingest KG features into machine learning applications due to computational efficiency and data sparsity challenges [149]. Hence, a common way of representing knowledge in a KG involves the use of embeddings. Aimed at encoding the latent properties of the graph in a continuous space, KG embeddings are low-dimensional dense representations of a given entity in a KG and its relationships [150]. Because latent semantics of the network are preserved while generating the embeddings, similarity in the embedding space implies the similarity of both concepts and relationships in the corresponding KG.

Although KG embeddings are not directly explainable, KGs can open the door to explainable affective computing via the integration of symbolic AI into subsymbolic AI approaches. These two alternative paradigms to AI offer different ways to model reality: on the one hand, subsymbolic AI does not make use of explicit symbols or rules to attain a model and relies on mechanisms to extract patterns from past experiences, such as statistical learning or DL methods. Because the model is not explicitly defined, these learning mechanisms can often lead to black-box models with low interpretability. On the other hand, symbolic AI aims to build a model via rules or formulas that can not only logically represent reality, leveraging objects and logical relationships in symbolic statements that are often hardcoded, but also be learned [151]. Because symbolic AI builds models that are explicitly represented via rules derived from reasoning, outputs are inherently explainable and can be described in human language. Hence, in affective computing, these two approaches to AI can be combined to leverage both the data-driven expressive power of subsymbolic AI (e.g., DL models) and the interpretability of symbols and domain-based knowledge (e.g., *commonsense* reasoning encoded in KGs). In this context, KG embeddings can be used

to integrate semantic and domain background knowledge in a machine-readable format into DL architectures, helping the model disambiguate concepts when dealing with large amounts of data while providing more meaningful explanations [152].

Several attempts exist in the literature combining KGs and subsymbolic AI techniques. Starting with emotion recognition in conversations, Ghosal et al. [153] present COSMIC, a framework that incorporates different elements of *commonsense* knowledge, such as mental states, events, and causal relations, to predict utterance emotions in a conversation. The proposed framework consists of three main stages: the extraction of context-independent features using pretrained language models, the extraction of knowledge-based (*commonsense*) features with the *commonsense* transformer model COMET pretrained on several *commonsense* KGs, and the integration of both features into dedicated GRU cells to model speaker states and intent. This approach not only achieved new SOTA results for emotion recognition in conversations but also provided better qualitative interpretations of predictions via the incorporation of *commonsense* knowledge-based features.

Also, in the context of emotion recognition in conversations, Wang et al. [154] use KGs to integrate emotional causality reasoning into empathetic response generation, improving the explainability of both the emotion recognition and the utterance generation tasks. In their work, the user's emotional experience is represented by using a series of emotional causality graphs via multihop reasoning over *commonsense* KGs. These graphs are then used to build an embedding and generate an empathetic response. While most existing works focus on what the emotion is and ignore how the emotion is evoked, Wang et al. [154] show that the use of KGs not only helps with creating a context-aware embedding that can be exploited to synthesize an empathetic utterance but also contributes toward an improved understanding and interpretability of the system.

By integrating relevant domain context, KGs not only help a DL model achieve better performance via richer features and embeddings but can also enable effective disambiguation and noise handling. For instance, in the context of ABSA where different aspects and sentiments can be mixed in a long sentence, KGs can help provide explainable and accurate aspect terms. For instance, Zhao et al. [155] study the problem of explainable ABSA by incorporating external domain knowledge into a pretrained BERT language model, and Zhong et al. [156] introduce KGAN, a KG augmented network that captures sentiment feature representations from a temporal, syntax, and knowledge-based perspective. Using attention mechanisms when learning these three representations, weights can be inspected to understand which words were attended the most from each different perspective.

The combination of KGs and subsymbolic AI approaches is also at the core of Sentic Computing [157]. Defined as a multidisciplinary approach to sentiment analysis at the crossroads between affective computing and *commonsense* computing, several algorithms and resources proposed under this umbrella enable the incorporation of external knowledge into DL architectures [158]. For instance, *AffectiveSpace* [159], [160] and *SenticNet* [161] are KG resources built to represent affective *commonsense* reasoning. From the algorithm's perspective, Sentic LSTM incorporates *commonsense* knowledge of sentiments into the end-to-end training of an LSTM model: the core of the method, fully described in [162], relies on the integration of relevant AffectiveSpace embeddings into LSTM gate mechanisms to control the flow of word-level information through the LSTM cell. Finally, Sentic LDA integrates knowledge from *SenticNet* in the calculation of word distributions within the standard LDA algorithm, enabling the shift from syntax to semantics in ABSA [158], [163].

D. In-Model Methods' Conclusion

In-model XAI methods have the potential to improve both predictive performance and model transparency in affective computing. While the methods discussed in this section can help overcome key XAI challenges in the field, they also face several issues. Attention requires post hoc analysis, and when used in architectures incorporating hierarchical mechanisms or transformer-based strategies, it can lead to hundreds of millions of parameters, requiring high computational power to be trained and further processing to produce a useful explanation. Generative models can be interpreted directly, but the discovery of factors requires expert knowledge, which can be impractical for some tasks. In addition, if too many factors are added to improve model capacity, a generative model can also become difficult to interpret. When the task at hand benefits from explicitly modeling relational structures in the data, graph representations can benefit both interpretability and model performance. However, there are many works in the literature incorporating graphs that are solely focused on improving predictive performance, adding additional complexity layers and parameters to be trained, and preventing effective interpretation when the graphs are not sparse. Finally, KGs can bridge the gap between data-driven analyses and human meaning, providing strong priors and exposing frameworks for interpretation. However, complex sets of rules can also be hard to interpret.

As we will discuss later in Section VI, any in-model XAI method has the potential to generate uninterpretable complex sets of patterns, which may need to be preprocessed further to avoid increasing the cognitive load of explainees [42]. In this way, a holistic approach to in-model XAI could be the key to unlock explainability in affective computing, with hybrid methods marrying the strengths of different techniques.

V. Postmodel XAI Methods for Affective Computing

The goal of postmodel XAI techniques is to understand how model outputs relate to inputs after model training. These techniques estimate feature importance by treating the model as a black box (model agnostic methods) or exploiting some aspect of the model's internal structure (e.g., activations of neurons in a neural network).

When the scope is local and the goal is to explain a given prediction, model agnostic methods approximate model

TABLE IV
SUMMARY OF POSTMODEL XAI METHODS FOR AFFECTIVE COMPUTING TASKS

Post-Model XAI Methods for Affective Computing			
Methodology	**Task**	**Reference**	**Multimodal**
Post-hoc XAI applied on discrete sets of inputs			
DeepLift	Pain Estimation	Liu et al., 2017 [172]	✗
CNN-based method	Depression Recognition	Zhou et al., 2018 [174]	✗
Shapley Values	Facial Emotion Recognition	Malik et al., 2021 [175]	✗
	EEG Emotion Recognition	Liew et al., 2021 [179]	✗
LRP	Driver Behaviour - Eye Gaze	Hwu et al., 2021 [177]	✗
	Pain Estimation	Prajod et al., 2022 [173]	✗
	Facial Emotion Recognition	Zhu et al., 2022 [176]	✗
LIME	Emotion Recognition	Heimerl et al., 2020 [178]	✗
		Chowdhury et al., 2021 [180]	✗
DIME	Not applied yet in affective computing	Lyu et al., 2022 [182]	✓
Multiviz	Emotion Recognition	Liang et al., 2022 [183]	✓
Post-hoc XAI applied on discrete sets of features			
DeepLift	Speech Emotion Recognition	Das et al., 2021 [186]	✗
Shapley Values	Personality Trait Recognition	Mehta et al., 2020 [190]	✗
	Mental wellbeing	Nishimura et al., 2022 [188]	✓
		LEWIS et al., 2021 [189]	✓
M2Lens	Sentiment Analysis	Wang et al., 2021 [191]	✓

computations by applying perturbations on inputs and observing the changes in model outputs. This is the case with Shapley values [164], where feature importance is approximated by averaging the marginal contribution of a given feature on model outputs for all possible permutations of other features [165], [166], and with LIME [167], where random perturbations are applied to inputs and locally weighted regression is used to identify the perturbations leading to the biggest change in model output. Local postmodel methods can also make use of the model's inner workings to approximate the relationship between inputs and outputs: for example, gradient-based approaches use the gradient of the output for a given class with respect to some inputs to quantify feature importance. This is the case with methods such as SmoothGrad [168], GradCAM [169], LRP [170], or DeepLift [171]. When the scope is global and the goal is to understand how the model works on average considering the whole dataset, explanations are generally obtained by either aggregating local explanations (e.g., using LIME or Shapley) or by applying a global surrogate model. The latter, also called an *explainer*, is by design a simpler model, such as a tree-based method, which approximates inputs and outputs of a more complex, black-box, model.

A. Postmodel XAI Applied on Discrete Sets of Unimodal Inputs

A major challenge facing postmodel methods in affective computing relates to representation learning and the requirement to have a discrete, interpretable, set of features to apply these methods. From a computational perspective, because bigger sets of features can lead to bottlenecks and heavy workloads in the computation of explanations, smaller sets of features are beneficial. From a modeling perspective, discrete sets of features can lead to high variance across samples, limiting the ability to generalize in different contexts (e.g., different languages). Because these handcrafted features rely on human perception, by design, they cannot encompass all the relevant variance in the data.

Hence, in affective computing, postmodel XAI techniques have been mainly applied on tasks where a single modality is considered and a discrete set of inputs, such as pixels or words, is used to learn data representations (see Table IV). For instance, in vision, postmodel XAI applications include pain estimation from facial expressions using DeepLift [172] and LRP [173], interpretable depression recognition from facial images using a CNN-based postmodel method [174], facial emotion recognition using an extension of Shapley values [175] and LRP [176], and prediction of driving behavior with LRP [177]. Using LIME on video frames for emotion recognition, Heimerl et al. [178] introduced NOVA, an annotation tool for emotional behavior analysis implementing a workflow that interactively incorporates the "human in the loop." In their work, the authors investigate how LIME can better assist nonexperts in terms of the trust, perceived self-efficacy, cognitive workload, and in creating the correct mental models about the system. With a user study of 53 participants, the *causability* and value of the proposed explanations are assessed, with results indicating that LIME can help users better understand the system compared to not having any explanation.

In EEG signal-based emotion recognition, postmodel XAI methods, such as Shapley values, have also been applied. In [179], for example, fuzzy ART (FA) techniques and genetic algorithms are used to amplify the signal-to-noise ratio and create clusters of features that are then fed into boosted decision trees. The importance of these clusters is interpreted via SHAP values, leading to insights on the contributions of both individual features and feature interactions toward predicting human effect.

In text-based sentiment analysis, LIME has been used to identify the most influential words in a sentence [180], and

more generally, in NLP, local and global postmodel explanations are also part of the Language Interpretability Tool [181], an open-source platform for visualization and understanding of NLP models.

In multimodal settings, Lyu et al. [182] propose DIME, an extension of LIME disentangling a multimodal model into unimodal contributions and multimodal interactions after training. While the experiments performed only involve the visual and text modalities due to the high computational cost of the method, user evaluations show that DIME can help researchers determine which unimodal or multimodal contributions are the dominant factors behind the model's prediction. An improvement of DIME aimed at improving its scalability is introduced in MULTIVIZ [183], a tool for analyzing the behavior of multimodal models that scaffold the problem of interpretability into unimodal importance, cross-modal interactions, multimodal representations, and multimodal predictions.

B. Postmodel XAI Applied on Handcrafted and Low-Level Features

Apart from being applied to pixels or words, postmodel XAI can also be applied to handcrafted features. While the inevitable information loss of manual features has propelled the development of other representation learning approaches [184], in certain tasks, discrete sets of representations derived from science can still be beneficial, especially when big pools of data are not available and when understanding the association between specific attributes is required [185]. In [186], for example, using a set of 88 handcrafted features and DeepLift, Das et al. investigate the ability of unsupervised learning methods to learn lower dimensional representations that can generalize over different languages. In affective computing for mental well-being, Alghowinem et al. [187] present a feature selection framework to automatically identify the most discriminative handcrafted features for depression detection, enabling interpretability while providing higher predictive accuracy. Using also handcrafted features to understand the psychological state of a given individual, in [188], Shapley values are used to assess the influence of sensor, behavioral, and weather factors in the psychological indexes of office workers. Similarly, in [189], behavioral insights are generated using Shapley values and packaged into personalized interventions, helping individuals build mindfulness habits.

When incorporating data-driven feature learning methods, learned features might not be directly interpretable. In these scenarios, postmodel methods have been applied in two different ways. The first way involves a hybrid approach that relies on bottom-up feature generation as part of the DL process and top-down integration of theory-based features. For instance, in personality prediction from text, Mehta et al. [190] propose a model integrating traditional psycholinguistic features with language embeddings, using Shapley values to understand the influence of psycholinguistic predictors in a particular personality trait prediction. While the interpretation of the model does not guarantee 100% coverage or faithfulness since language embeddings are not considered when computing Shapley values, the method generates insights into how different psycholinguistic features influence model outcomes.

A second way of generating post hoc explanations in models reliant on data-driven representations involves the use of a discrete set of handcrafted features relevant to the task at hand, but not necessarily used in the modeling process. In M2-Lens [191], this approach is followed to explain multimodal models for sentiment analysis. Providing explanations on intramodal and intermodal interactions at different levels, M2Lens uses a set of handcrafted features for each modality and relies on Shapley values to compute the importance of each modality. Although explaining model performance using a finite set of features can provide insights into how a given model is influenced by those specific features, this approach can only provide a relative view of feature importance. While this can be useful in tasks where a defined set of interpretable features is relevant for decision-making, it can be misleading for researchers trying to understand how their model works. For instance, one expert in [191] concludes that an LSTM model does not rely on text to learn sentiment when, in fact, the conclusion needs to be framed in relative terms: because part-of-speech is the only text feature being considered to quantify the importance of the text modality, the conclusion should be that this specific feature is not as relevant as other features being considered in the computation of importances. We cannot infer that the text modality on its own is not relevant nor the model is unable to learn sentiment in text: if, in fact, the model is using only POS tags as the text feature, we may be constraining its ability to extract information and variance from the language modality by only feeding this manual feature. If, on the other hand, the model is using other text features such as word embeddings, then, because we are not considering these language embeddings in the computation of feature importances, we cannot conclude either that the model is not paying attention to text to generate its predictions.

C. Postmodel Methods' Conclusion

Postmodel XAI techniques have been proved effective in unimodal cases and modeling scenarios involving finite sets of handcrafted features. However, in multimodal tasks, postmodel methods need to be adapted to appropriately account for intramodal and cross-modal interactions, usually leading to high computational costs. Depending on the goal of the explanations and who are the explainees, a discrete set of manual features or highlighted inputs can be beneficial (e.g., NOVA [178]). However, when the goal is to understand how the model operates in order to improve it, faithfulness is crucial to avoid overtrusting or undertrusting the system. In this context, simplistic explanations can lead to disregarding important aspects of the model influencing its performance.

VI. Discussion and Future Directions

We have reviewed and categorized several approaches to explain the output of multimodal affective computing systems. As evidenced by these publications, recent advances in XAI have made it possible to explain the output of affective

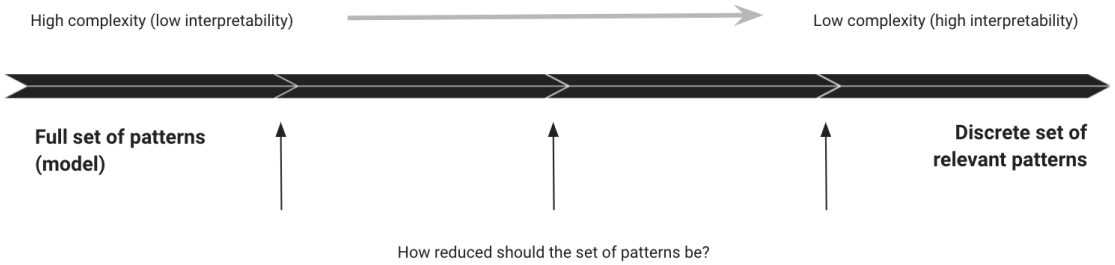

Fig. 1. XAI understood as a dimensionality reduction problem: explanation goals and explainee need to determine the size of the required subset of patterns composing the explanation.

computing systems without negatively impacting predictive performance. The field, however, is still in its infancy, and several challenges still lie ahead preventing transparent affective computing systems from scaling. In the following, we discuss the fundamental difficulties to be addressed by researchers and directions for future research.

A. Dimensionality Reduction as an Analogy for Generating Explanations

The ultimate goal of XAI in affective computing is extracting a relevant subset of patterns that explain the relationship between model inputs and outputs to help a user better understand the behavior of an affective system. In this way, the complete set of patterns being captured from the data is given by the model, which is just a function mapping a set of inputs to a given output or target of interest. A model can be given by a neural network architecture, a differential equation system, an exhaustive set of deterministic rules or heuristics, or a set of ontologies and knowledge-based logic laws. Whatever its form, a model needs to have enough expressive power and capacity to model high-dimensional and sparse data. In the case of affective computing, where accounting for the intertwined relationships between different modalities and context factors is crucial, to achieve high performance and be able to exploit the patterns in the data, models necessarily need to have great flexibility and expressivity. Thus, an affective computing system naturally involves complex models since it is the nature of the task itself that determines the complexity of the full set of patterns that need to be considered.

The problem of explaining the output of affective computing systems can be therefore reframed as a function approximation problem: given a complex set of patterns considered by a model, how can we extract a relevant subset of patterns to help a user understand the relationship between model inputs and outputs? (See Fig. 1). From this perspective, generating explanations is no different from dimensionality reduction in statistical learning and machine learning. Hence, both to understand affective data, high dimensional and unstructured, and to understand how an affective model reasons and operates, we need to reduce a set of complex patterns to a relevant summary to be consumed by a human, minimizing cognitive load and dilution of the most important information. As discussed in Section III, the key challenge with the former is the lack of standardization methods to evaluate, explore, and compare multimodal datasets. The key challenge with the latter, however, is how to ensure that the key information is preserved, and the final set of relationships to be considered is representative of the full set of patterns exploited by the model.

While we acknowledge that not all applications need complete faithfulness and explanations can still be useful even when the subset of patterns does not have full coverage of the real patterns leveraged by the model, faithful representations of model patterns are usually required when building explanations for researchers in the field, who need a deep understanding of model operations in order to improve them, as well as decision-makers in high-stake scenarios, such as healthcare or education, where model reliability is critical. Considering the complexity of the full set of patterns as a given due to the complexity of affective computing tasks, further research is needed on how to reduce this high-dimensional set of patterns into a manageable, human-friendly, summary of relationships while maintaining the variance and key properties of the data in the transformation.

B. Isolated Research Efforts Can Lead to Shifting the Burden of XAI

Because a model is just a function relating inputs to outputs and affective computing tasks are complex, involving high-dimensional data, as well as multimodal and time dependencies, we argue that any model with enough capacity will generate a complex set of patterns or rules. In this way, the model itself can be thought of as a very complex explanation with low interpretability (see Fig. 1).

While some in-model XAI methods are claimed to be more interpretable by design, such as theory-driven generative models or knowledge-based algorithms, a complex task can require hundreds of factors to be considered in a generative model or thousands of ontologies in a knowledge-based method. In this way, high-complexity tasks can lead to generative models that are not directly interpretable or large collections of intertwined knowledge-based rules, which would need to be mined. Hence, moving from one modeling framework to another, such as from discriminative models to generative models, will not solve the problem of generating an explanation: because the task in itself is complex, the set of rules will still be complex,

and we are just shifting the XAI burden from one framework to another.

C. No Free Lunch: Explanation Goals and User Needs Matter

Assuming that we have a continuous space between a very complex set of model patterns and a very simple relevant subset of those patterns (see Fig. 1), where do we stop? How reduced should the subset be? The type of application, the users involved, the goal of the explanations, and even the frequency of explanations will determine the capacity of the explainee to absorb information and the reduction that needs to take place during the transformation. For instance, when considering a researcher as the final user of the XAI output and assuming that the goal of the explanation is to provide enough information about the model for the researcher to improve it, because this user is already knowledgeable about how the model works, in this scenario, we would expect a higher capacity to absorb information. Hence, more complex summaries guaranteeing faithfulness to the system's decision-making would be beneficial. On the other hand, when considering laypeople using an intelligent tutoring system to learn how to best behave in a job interview, the explanation exposed needs to be very simple, yet actionable. Assuming that, in this case, the users require explanations to understand why their score in the learning system is low and how to adapt their behavior to score higher in a job interview, the output summary needs to be targeted to the specific aspects that led to the low score in order to help the users focus their learning on what is important.

While these are hypothetical scenarios where we assume certain user needs, user requirements should not be assumed without evaluation. In fact, before even considering XAI techniques or methods, any affective computing application should consider why an explanation is needed in the first place. If user needs are not properly understood, then the extent to which an explanation should be simplified is unknown. Because models are simplifications of reality and explanations are simplifications of the models underneath, oversimplification and incompleteness can lead to incorrect interpretations of explanations, limiting their usefulness in practice. In the same way, depending on who are the explainees, an explanation that is complete and complex can also lead to wrong conclusions and/or to being dismissed. This misalignment between expected interpretations and actual understanding of explanations can limit practical correctness in the same way that flawed XAI algorithms can limit technical correctness. While quantitative analyses can help mitigate the risks of technical incorrectness, to solve for practical correctness careful user design is needed.

In this review, we have seen that almost no research to date includes user evaluations or assessments of the usefulness of the explanations, that is, the *causability* of the XAI outputs, when, in fact, the application itself and the user needs should determine the choice of the XAI method, the scope of the explanation, and the extent of the dimensionality reduction applied on the model patterns. While we believe that this is the case because explanation evaluations are not standardized yet in the field, the human dimension of explanations should not be ignored nor avoided.

D. XAI and Causality

Essentially, an explanation can be thought of as an answer to a *why-question*. In this review, we have reduced the scope to consider only associative reasoning, considering an explanation as a subset of relevant patterns giving information on what inputs relate to outputs and which part(s) of the input and/or model structure influence its final recommendations. Understanding feature importance and the relationship between model inputs and outputs is, however, just the beginning of the XAI journey, and hence, this article has focused on gathering evidence to understand whether this first stage is even possible in affective computing. Nevertheless, when considering again the case of an intelligent tutoring system, how can we guarantee that, if the user changes the specific traits leading to a low score, no other traits will be changed, and therefore, the final score will be higher? While out of scope in this review, the next stage in the XAI journey would involve interventionist and counterfactuals reasoning, that is, the ability to understand not only which inputs relate to model outputs but also which features, traits, or behaviors need to be changed and by how much to flip the output of an affective computing system. In affective computing, this type of reasoning answers the question of not only *why* but also *what if* can be useful in a wide spectrum of applications, ranging from learning systems to diagnostic tools in healthcare.

Closely related to counterfactuals are the topics of causality and causal modeling. Because counterfactuals involve reasoning about why things happened and imagining the consequences of different actions in hindsight [192], adding this additional layer of reasoning into explanations requires models to identify and extract causal patterns from the data. While causal modeling is outside of the scope of this article, we believe that causal inference can be a powerful modeling tool for future explainable affective computing, moving the field forward toward more valuable and richer explanations aligned with user needs.

E. Word of Caution: Explainable Affective Computing Does Not Equate to Ethical Affective Computing

XAI does not justify illegitimate use cases of affective computing. A model can be explainable and still lead to unacceptable risks. Because AI models are uncertain and no AI model is perfect, impact assessments [193] need to be carried out in order to evaluate potential risks and benefits for all involved stakeholders. When potential risks outweigh expected benefits, different approaches other than AI should be sought. Furthermore, XAI is also only one key aspect of transparency, the latter involving responsible communication at all levels, including system limitations, scope, and consent among other aspects [23]. At the same time, transparency is only one component of responsible AI [78] and, by itself, does not guarantee that the developed system will respond to responsible AI principles in practice. For these reasons,

premodel, in-model, and postmodel explainability should not be seen as sufficient prechecks to release affective systems into real-world applications, but rather, as just a component in a broader picture, including other applicable transparency, privacy, accountability, and reliability goals.

Explanations do not always lead to positive effects either. In fact, explanations can intentionally be used to deceive and manipulate people [194], as well as unintentionally generate negative consequences downstream [195]. In affective computing, where modeling is primarily focused on human behavior, both manipulative explanations and pitfalls can lead to emotional damage and infringement on human rights. Moreover, inducing trust and overreliance on systems that should not be trusted can negatively affect individuals at scale. In this context, we advocate for a human-centered approach to XAI, starting by questioning the need for explanations and carefully assessing who can benefit from them. Rather than prescriptive, we see explanations as prompts for further thinking, encouraging reflection rather than automatic acceptance. From this perspective, explanations can be a tool aimed at giving agency to people, helping end users improve their own capabilities rather than replacing their judgment. While XAI has been primarily focused on opening the black box of DL to help researchers improve algorithmic performance, future research requires multidisciplinary collaboration and cross-disciplinary knowledge exchange, involving not only the development of techniques for algorithmic transparency but also the practice of responsible and human-centered XAI design.

VII. Conclusion

Affective computing is an active and challenging multidisciplinary research area with unprecedented potential to change the way humans interact with technology. While the last decades have witnessed vast progress in the field, the black-box nature of affective computing systems is one of the key aspects preventing them from scaling into real-world applications. In this article, we have explored whether this issue can be overcome with no extra cost to practical predictive performance. Toward this end, we have analyzed and categorized examples of work implementing XAI methods in affective computing applications and related tasks, with a focus on those implementations achieving comparable results to the SOTA or even improving them. While explainable affective computing is still nascent and several challenges still are ahead, we conclude that first-stage explanations relating inputs to outputs are already technically feasible and do not necessarily involve the use of simplistic modeling methods unable to capture the complexity of affective computing tasks. Future research needs to address aspects such as faithfulness and *causability*, but the path toward explainable and transparent affective computing systems has already been cleared.

References

[1] J. Lear, *Aristotle: The Desire to Understand*. Cambridge, U.K.: Cambridge Univ. Press, 1988.

[2] J. Han, Z. Zhang, M. Pantic, and B. Schuller, "Internet of Emotional People: Towards continual affective computing cross cultures via audiovisual signals," *Future Gener. Comput. Syst.*, vol. 114, pp. 294–306, Jan. 2021.

[3] M. Conti, A. Passarella, and S. K. Das, "The Internet of People (IoP): A new wave in pervasive mobile computing," *Pervasive Mobile Comput.*, vol. 41, pp. 1–27, Oct. 2017.

[4] J. Miranda et al., "From the Internet of Things to the Internet of People," *IEEE Internet Comput.*, vol. 19, no. 2, pp. 40–47, Mar. 2015.

[5] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017.

[6] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. M. Rahaman, and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," *J. Netw. Comput. Appl.*, vol. 149, Jan. 2020, Art. no. 102447.

[7] C. R. Fetsch, A. Pouget, G. C. DeAngelis, and D. E. Angelaki, "Neural correlates of reliability-based cue weighting during multisensory integration," *Nature Neurosci.*, vol. 15, no. 1, pp. 146–154, Jan. 2012.

[8] M. L. Dumitru, P. Achille, and M. Andriy, *Multisensory Integration: Brain, Body and the World*. Lausanne, Switzerland: Frontiers Media, 2016.

[9] Y.-H.-H. Tsai, M. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, p. 1823.

[10] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[11] P. C. Ellsworth and K. R. Scherer. *Appraisal Processes in Emotion*. New York, NY, USA: Oxford Univ. Press, 2003.

[12] D. C. Ong et al., "Modeling emotion in complex stories: The Stanford emotional narratives dataset," *IEEE Trans. Affect. Comput.*, vol. 12, no. 3, pp. 579–594, Jul. 2021.

[13] K. Hoemann et al., "Context-aware experience sampling reveals the scale of variation in affective experience," *Sci. Rep.*, vol. 10, no. 1, pp. 1–16, Jul. 2020.

[14] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychol. Sci. Public Interest*, vol. 20, no. 1, pp. 1–68, Jul. 2019.

[15] A. Heimerl, T. Baur, and E. André, "A transparent framework towards the context-sensitive recognition of conversational engagement," in *Proc. MRC ECAI*, 2020, pp. 7–16.

[16] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.

[17] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001.

[18] T. Xu, W. Jennifer, K. Sinan, and G. Hatice, "Investigating bias and fairness in facial expression recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 506–523.

[19] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *Proc. IEEE Workshop Adv. Robot. Social Impacts (ARSO)*, Mar. 2017, pp. 1–7.

[20] J. H. Shen, "Affective computing and crowdsourcing: Subjective labels and sequential effects," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2019.

[21] R. W. Picard, "Affective computing: Challenges," *Int. J. Hum.-Comput. Stud.*, vol. 59, nos. 1–2, pp. 55–64, 2003.

[22] R. Cowie, "Ethical issues in affective computing," in *The Oxford Handbook of Affective Computing*. New York, NY, USA: Oxford Univ. Press, 2015, pp. 334–348.

[23] J. Hernandez et al., "Guidelines for assessing and minimizing risks of emotion recognition applications," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2021, pp. 1–8.

[24] M. B. Ammar, M. Neji, A. M. Alimi, and G. Gouardères, "The affective tutoring system," *Exp. Syst. Appl.*, vol. 37, no. 4, pp. 3013–3023, 2010.

[25] S. Greene, H. Thapliyal, and A. Caban-Holt, "A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health," *IEEE Consum. Electron. Mag.*, vol. 5, no. 4, pp. 44–56, Oct. 2016.

[26] C. Zucco, B. Calabrese, and M. Cannataro, "Sentiment analysis and affective computing for depression monitoring," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 1988–1995.

[27] C. Conati, K. Porayska-Pomsta, and M. Mavrikis, "AI in education needs interpretable machine learning: Lessons from open learner modelling," 2018, *arXiv:1807.00154*.

[28] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. R. Müller, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700. Cham, Switzerland: Springer, 2019.

[29] G. Joshi, R. Walambe, and K. Kotecha, "A review on explainability in multimodal deep neural nets," *IEEE Access*, vol. 9, pp. 59800–59821, 2021.

[30] A. L. Hunkenschroer and A. Kriebitz, "Is AI recruiting (un)ethical? A human rights perspective on the use of AI for hiring," *AI Ethics*, vol. 3, no. 1, pp. 199–213, Feb. 2023.

[31] S. T. Mueller, R. R. Hoffman, W. Clancey, A. Emrey, and G. Klein, "Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI," 2019, *arXiv:1902.01876*.

[32] H. J. Escalante et al., "Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos," 2018, *arXiv:1802.00745*.

[33] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello, "Bias and fairness in multimodal machine learning: A case study of automated video interviews," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 268–277.

[34] D. V. Carvalho, M. E. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, Jul. 2019.

[35] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.* Cham, Switzerland: Springer, 2019, pp. 563–574.

[36] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[37] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[38] W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," *Knowl.-Based Syst.*, vol. 263, Mar. 2023, Art. no. 110273.

[39] R. Dwivedi et al., "Explainable AI (XAI): Core ideas, techniques, and solutions," *ACM Comput. Surveys*, vol. 55, no. 9, pp. 1–33, Sep. 2023.

[40] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[41] A. Holzinger, B. Malle, A. Saranti, and B. Pfeifer, "Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI," *Inf. Fusion*, vol. 71, pp. 28–37, Jul. 2021.

[42] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2018.

[43] J. Pearl and M. Dana, *The Book of Why: The New Science of Cause and Effect*. New York, NY, USA: Basic books, 2018.

[44] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statist. Surveys*, vol. 16, pp. 1–85, Jan. 2022.

[45] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning—A brief history, state-of-the-art and challenges," in *Proc. ECML PKDD Workshops, Eur. Conf. Mach. Learn. Knowl. Discovery Databases (ECML PKDD)*. Cham, Switzerland: Springer, Sep. 2020, pp. 417–431.

[46] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019.

[47] R. Marcinkevičs and J. E. Vogt, "Interpretability and explainability: A machine learning zoo mini-tour," 2020, *arXiv:2012.01805*.

[48] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 4, p. e1312, Jul. 2019.

[49] A. Vinciarelli et al., "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 69–87, Jan. 2012.

[50] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, Nov. 2009.

[51] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 273–291, Jul. 2014.

[52] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," 2016, *arXiv:1609.01885*.

[53] D. Dresvyanskiy, M. Wolfgang, and K. Alexey, "Deep learning based engagement recognition in highly imbalanced data," in *Proc. Int. Conf. Speech Comput.* Cham, Switzerland: Springer, 2021, pp. 166–178.

[54] T. Lombrozo, "The structure and function of explanations," *Trends Cognit. Sci.*, vol. 10, no. 10, pp. 464–470, Oct. 2006.

[55] H. Zeng, X. Wang, A. Wu, Y. Wang, Q. Li, A. Endert, and H. Qu, "*EmoCo*: Visual analysis of emotion coherence in presentation videos," *IEEE Trans. Visualizat. Comput. Graph.*, vol. 26, no. 1, pp. 927–937, Aug. 2019.

[56] H. Zeng et al., "EmotionCues: Emotion-oriented visual summarization of classroom videos," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 7, pp. 3168–3181, Jul. 2021.

[57] X. Wang, Y. Ming, T. Wu, H. Zeng, Y. Wang, and H. Qu, "DeHumor: Visual analytics for decomposing humor," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 12, pp. 4609–4623, Dec. 2022.

[58] C. Wang, P. Lopes, T. Pun, and G. Chanel, "Towards a better gold standard: Denoising and modelling continuous emotion annotations based on feature agglomeration and outlier regularisation," in *Proc. Audio/Visual Emotion Challenge Workshop*, Oct. 2018, pp. 73–81.

[59] H. J. Escalante et al., "Modeling, recognizing, and explaining apparent personality from videos," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 894–911, Apr. 2022.

[60] M. F. Carbonell, M. Boman, and P. Laukka, "Comparing supervised and unsupervised approaches to multimodal emotion recognition," *PeerJ Comput. Sci.*, vol. 7, p. e804, Dec. 2021.

[61] H. Elhamdadi, S. Canavan, and P. Rosen, "AffectiveTDA: Using topological data analysis to improve analysis and explainability in affective computing," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 1, pp. 769–779, Sep. 2021.

[62] L. Chen, K. Wang, M. Li, M. Wu, W. Pedrycz, and K. Hirota, "K-means clustering-based kernel canonical correlation analysis for multimodal emotion recognition in human–robot interaction," *IEEE Trans. Ind. Electron.*, vol. 70, no. 1, pp. 1016–1024, Jan. 2022.

[63] N. Hajarolasvadi and H. Demirel, "3D CNN-based speech emotion recognition using K-means clustering and spectrograms," *Entropy*, vol. 21, no. 5, p. 479, May 2019.

[64] S. Khan, L. Chen, and H. Yan, "Co-clustering to reveal salient facial features for expression recognition," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 348–360, Apr. 2020.

[65] I. Siegert, R. Böck, and A. Wendemuth, "Using a PCA-based dataset similarity measure to improve cross-corpus emotion recognition," *Comput. Speech Lang.*, vol. 51, pp. 1–23, Sep. 2018.

[66] S. Kang, D. Kim, and Y. Kim, "A visual-physiology multimodal system for detecting outlier behavior of participants in a reality TV show," *Int. J. Distrib. Sensor Netw.*, vol. 15, no. 7, Jul. 2019, Art. no. 155014771986488.

[67] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello, "Integrating psychometrics and computing perspectives on bias and fairness in affective computing: A case study of automated video interviews," *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 84–95, Nov. 2021.

[68] J. Kossaifi et al., "SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 1022–1040, Mar. 2021.

[69] A. Zadeh et al., "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meet. Assoc. Comput. Linguistics (ACL)*, vol. 1, 2018, pp. 2236–2246.

[70] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–8.

[71] T. Gebru et al., "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021.

[72] M. Pushkarna, A. Zaldivar, and O. Kjartansson, "Data cards: Purposeful and transparent dataset documentation for responsible AI," 2022, *arXiv:2204.01075*.

[73] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, "The dataset nutrition label," *Data Protection and Privacy*, vol. 12, no. 12, p. 1, 2020.

[74] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," in *Proc. Trans. Assoc. Comput. Linguistics*, vol. 6, 2018, pp. 587–604.

[75] A. Tsymbal, "The problem of concept drift: Definitions and related work," Dept. Comput. Sci., Trinity College Dublin, Dublin, Ireland, Tech. Rep. 2, 2004, p. 58.

[76] L. P. Morency, S. Sakti, B. W. Schuller, and S. Ultes, "Multimodal machine learning for social interaction with ageing individuals," in *Multimodal Agents for Ageing and Multicultural Societies: Communications of NII Shonan Meetings*. Singapore: Springer, 2021, pp. 61–70.

[77] P. P. Liang et al., "MultiBench: Multiscale benchmarks for multimodal representation learning," 2021, *arXiv:2107.07502*.

[78] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[79] U. Ehsan, Q. V. Liao, M. Müller, M. O. Riedl, and J. D. Weisz, "Expanding explainability: Towards social transparency in AI systems," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2021, pp. 1–19.

[80] U. Aïvodji, A. Bolot, and S. Gambs, "Model extraction from counterfactual explanations," 2020, *arXiv:2009.01884*.

[81] S. Goethals, K. Sörensen, and D. Martens, "The privacy issue of counterfactual explanations: Explanation linkage attacks," 2022, *arXiv:2210.12051*.

[82] K. Bonawitz et al., "Towards federated learning at scale: System design," in *Proc. Mach. Learn. Syst.*, vol. 1, 2019, pp. 374–388.

[83] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 28, 2022, doi: 10.1109/TNNLS.2022.3160699.

[84] Y. S. Can and C. Ersoy, "Privacy-preserving federated deep learning for wearable IoT-based biomedical monitoring," *ACM Trans. Internet Technol.*, vol. 21, no. 1, pp. 1–17, Feb. 2021.

[85] O. Rudovic et al., "Personalized federated deep learning for pain estimation from face images," 2021, *arXiv:2101.04800*.

[86] P. Pu Liang et al., "Think locally, act globally: Federated learning with local and global representations," 2020, *arXiv:2001.01523*.

[87] P. Voigt and A. Von Dem Bussche, *The EU General Data Protection Regulation (GDPR)*, vol. 10, 1st ed. Cham, Switzerland: Springer, 2017, Art. no. 3152676.

[88] G. Haddon-Hill, K. Kusumam, and M. Valstar, "A simple baseline for evaluating expression transfer and anonymisation in video transfer," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Sep. 2021, pp. 01–08.

[89] T. Feng and S. Narayanan, "Privacy and utility preserving data transformation for speech emotion recognition," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2021, pp. 1–7.

[90] P. Pu Liang et al., "Learning language and multimodal privacy-preserving markers of mood from mobile data," 2021, *arXiv:2106.13213*.

[91] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–15.

[92] S. Jain and B. C. Wallace, "Attention is not explanation," 2019, *arXiv:1902.10186*.

[93] D. Pruthi, M. Gupta, B. Dhingra, G. Neubig, and Z. C. Lipton, "Learning to deceive with attention-based explanations," 2019, *arXiv:1909.07913*.

[94] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," 2019, *arXiv:1908.04626*.

[95] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, p. 6558.

[96] Y. Wang, J. Wu, and K. Hoashi, "Multi-attention fusion network for video-based emotion recognition," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 595–601.

[97] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Inf. Fusion*, vol. 68, pp. 46–53, Apr. 2021.

[98] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, p. 2225.

[99] Y. Gu et al., "Human conversation analysis using attentive multimodal networks with hierarchical encoder–decoder," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 537–545.

[100] U. Ahmed, R. H. Jhaveri, G. Srivastava, and J. C.-W. Lin, "Explainable deep attention active learning for sentimental analytics of mental disorder," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, Jul. 2022.

[101] S. Han, R. Mao, and E. Cambria, "Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings," 2022, *arXiv:2209.07494*.

[102] L. Hemamou, A. Guillon, J.-C. Martin, and C. Clavel, "Multimodal hierarchical attention neural network: Looking for candidates behaviour which impact Recruiter's decision," *IEEE Trans. Affect. Comput.*, early access, Sep. 16, 2021, doi: 10.1109/TAFFC.2021.3113159.

[103] T.-S. Nguyen, Z. Wu, and D. C. Ong, "Attention uncovers task-relevant semantics in emotional narrative understanding," *Knowl.-Based Syst.*, vol. 226, Aug. 2021, Art. no. 107162.

[104] D. Rodrigues, N. Kreif, A. Lawrence-Jones, M. Barahona, and E. Mayer, "Reflection on modern methods: Constructing directed acyclic graphs (DAGs) with domain experts for health services research," *Int. J. Epidemiol.*, vol. 51, no. 4, pp. 1339–1348, Aug. 2022.

[105] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[106] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr. 2003, pp. 1–4.

[107] M. Song, C. Chen, and M. You, "Audio-visual based emotion recognition using tripled hidden Markov model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, pp. 1–4.

[108] D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Ganzalez, and H. Sahli, "Audio visual emotion recognition based on triple-stream dynamic Bayesian network models," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* Berlin, Germany: Springer, 2011, pp. 609–618.

[109] I. Cohen, N. Sebe, F. G. Gozman, M. C. Cirelo, and T. S. Huang, "Learning Bayesian network classifiers for facial expression recognition both labeled and unlabeled data," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2003, pp. 1–7.

[110] H. Lee, Y. Sang Choi, S. Lee, and I. P. Park, "Towards unobtrusive emotion recognition for affective social communication," in *Proc. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2012, pp. 260–264.

[111] M. M. H. El Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using Gaussian mixture vector autoregressive models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, pp. 1–4.

[112] A. Metallinou, S. Lee, and S. Narayanan, "Audio-visual emotion recognition using Gaussian mixture models for face and voice," in *Proc. 10th IEEE Int. Symp. Multimedia*, Dec. 2008, pp. 250–257.

[113] L. Pang and C.-W. Ngo, "Mutlimodal learning with deep Boltzmann machine for emotion prediction in user generated videos," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, Jun. 2015, pp. 619–622.

[114] T. Horii, Y. Nagai, and M. Asada, "Modeling development of multimodal emotion perception guided by tactile dominance and perceptual improvement," *IEEE Trans. Cognit. Develop. Syst.* vol. 10, no. 3, pp. 762–775, Feb. 2018.

[115] D. Koller and F. Nir, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[116] I. Perikos, S. Kardakis, and I. Hatzilygeroudis, "Sentiment analysis using novel and interpretable architectures of hidden Markov models," *Knowl.-Based Syst.*, vol. 229, Oct. 2021, Art. no. 107332.

[117] C.-Y. Ting, W.-N. Cheah, and C. C. Ho, "Student engagement modeling using Bayesian networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2013, pp. 2939–2944.

[118] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 5, pp. 867–881, Oct. 2010.

[119] C.-C. Lee, C. Busso, S. Lee, and S. S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2009, pp. 1–4.

[120] R. Krishnan, U. Shalit, and D. Sontag, "Structured inference networks for nonlinear state space models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017, pp. 1–9.

[121] L. Li et al., "Hybrid deep neural network-hidden Markov model (DNN-HMM) based speech emotion recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 312–317.

[122] T. Zhi-Xuan, H. Soh, and D. Ong, "Factorized inference in deep Markov models for incomplete multimodal time series," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 6, 2020, pp. 10334–10341.

[123] Y.-H. Hubert Tsai, P. Pu Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," 2018, *arXiv:1806.06176*.

[124] D. C. Ong, H. Soh, J. Zaki, and N. D. Goodman, "Applying probabilistic programming to affective computing," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 306–317, Apr. 2021.

[125] J.-W. Van De Meent, B. Paige, H. Yang, and F. Wood, "An introduction to probabilistic programming," 2018, *arXiv:1809.10756*.

[126] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A survey on knowledge graphs: Representation, acquisition, and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, Feb. 2022.

[127] J. Zhou, X. Zhang, Y. Liu, and X. Lan, "Facial expression recognition using spatial–temporal semantic graph network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1961–1965.

[128] Y. Liu, X. Zhang, J. Zhou, and L. Fu, "SG-DSN: A semantic graph-based dual-stream network for facial expression recognition," *Neurocomputing*, vol. 462, pp. 320–330, Oct. 2021.

[129] E. Ghaleb, A. Mertens, S. Asteriadis, and G. Weiss, "Skeleton-based explainable bodily expressed emotion recognition through graph convolutional networks," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 1–8.

[130] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.

[131] M. Ye, C. L. P. Chen, and T. Zhang, "Hierarchical dynamic graph convolutional network with interpretability for EEG-based emotion recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 9, 2022, doi: 10.1109/TNNLS.2022.3225855.

[132] C. Li et al., "Effective emotion recognition by learning discriminative graph topologies in EEG brain networks," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 2, 2023, doi: 10.1109/TNNLS.2023.3238519.

[133] C. Zhang, Q. Li, and D. Song, "Aspect-based sentiment classification with aspect-specific graph convolutional networks," 2019, *arXiv:1909.03477*.

[134] X. Hou et al., "Graph ensemble learning over multiple dependency trees for aspect-level sentiment classification," 2021, *arXiv:2103.11794*.

[135] S. Pang, Y. Xue, Z. Yan, W. Huang, and J. Feng, "Dynamic and multi-channel graph convolutional networks for aspect-based sentiment analysis," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP*. Association for Computational Linguistics, 2021, pp. 2627–2636.

[136] R. Li, H. Chen, F. Feng, Z. Ma, X. Wang, and E. Hovy, "DualGCN: Exploring syntactic and semantic information for aspect-based sentiment analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Nov. 14, 2022, doi: 10.1109/TNNLS.2022.3219615.

[137] T. He and X. Jin, "Image emotion distribution learning with graph convolutional networks," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 382–390.

[138] P. Xu, Z. Liu, G. I. Winata, Z. Lin, and P. Fung, "EmoGraph: Capturing emotion correlations using graph networks," 2020, *arXiv:2008.09378*.

[139] M. Zhang, Y. Liang, and H. Ma, "Context-aware affective graph reasoning for emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 151–156.

[140] S. Wu, L. Zhou, Z. Hu, and J. Liu, "Hierarchical context-based emotion recognition with scene graphs," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 26, 2022, doi: 10.1109/TNNLS.2022.3196831.

[141] Q. Gao, H. Zeng, G. Li, and T. Tong, "Graph reasoning-based emotion recognition network," *IEEE Access*, vol. 9, pp. 6488–6497, 2021.

[142] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," 2019, *arXiv:1908.11540*.

[143] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, and R. Li, "Conversational emotion recognition using self-attention mechanisms and graph neural networks," in *Proc. INTERSPEECH*, Oct. 2020, pp. 2347–2351.

[144] T. Ishiwatari, Y. Yasuda, T. Miyazaki, and J. Goto, "Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 7360–7370.

[145] H. Xu, Z. Yuan, K. Zhao, Y. Xu, J. Zou, and K. Gao, "GAR-Net: A graph attention reasoning network for conversation understanding," *Knowl.-Based Syst.*, vol. 240, Mar. 2022, Art. no. 108055.

[146] R. Joshi, V. Balachandran, S. Vashishth, A. Black, and Y. Tsvetkov, "DialoGraph: Incorporating interpretable strategy-graph networks into negotiation dialogues," 2021, *arXiv:2106.00920*.

[147] J. Yang et al., "MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences," 2020, *arXiv:2010.11985*.

[148] M. Palmonari and P. Minervini, "Knowledge graph embeddings and explainable AI," in *Knowledge Graphs for Explainable Artificial Intelligence: Foundations, Applications and Challenges*, vol. 47. 2020, pp. 193–209.

[149] Y. Dai, S. Wang, N. N. Xiong, and W. Guo, "A survey on knowledge graph embedding: Approaches, applications and benchmarks," *Electronics*, vol. 9, no. 5, p. 750, 2020.

[150] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2724–2743, Dec. 2017.

[151] E. Ilkou and M. Koutraki, "Symbolic vs sub-symbolic AI methods: Friends or enemies?" in *Proc. CIKM Workshops*, 2020, pp. 1–8.

[152] I. Tiddi and S. Schlobach, "Knowledge graphs as tools for explainable machine learning: A survey," *Artif. Intell.*, vol. 302, Jan. 2022, Art. no. 103627.

[153] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "COSMIC: Commonsense knowledge for emotion identification in conversations," 2020, *arXiv:2010.02795*.

[154] J. Wang, W. Li, P. Lin, and F. Mu, "Empathetic response generation through graph-based multi-hop reasoning on emotional causality," *Knowl.-Based Syst.*, vol. 233, Dec. 2021, Art. no. 107547.

[155] A. Zhao and Y. Yu, "Knowledge-enabled BERT for aspect-based sentiment analysis," *Knowl.-Based Syst.*, vol. 227, Sep. 2021, Art. no. 107220.

[156] Q. Zhong, L. Ding, J. Liu, B. Du, H. Jin, and D. Tao, "Knowledge graph augmented network towards multiview representation learning for aspect-based sentiment analysis," 2022, *arXiv:2201.04831*.

[157] E. Cambria and A. Hussain, "Sentic computing," *Cognit. Comput.*, vol. 7, no. 2, pp. 183–185, Apr. 2015.

[158] Y. Susanto, E. Cambria, B. C. Ng, and A. Hussain, "Ten years of sentic computing," *Cognit. Comput.*, vol. 14, no. 1, pp. 5–23, Jan. 2022.

[159] E. Cambria, J. Fu, F. Bisio, and S. Poria, "AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 29, no. 1, 2015, pp. 1–7.

[160] E. Cambria and A. Hussain, "Sentic computing," *Marketing*, vol. 59, no. 2, pp. 557–577, 2012.

[161] E. Cambria, D. Olsher, and D. Rajagopal, "SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1–7.

[162] Y. Ma, H. Peng, and E. Cambria, "Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–8.

[163] S. Poria, I. Chaturvedi, E. Cambria, and F. Bisio, "Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 4465–4473.

[164] H. W. Kuhn and W. T. Albert, *Contributions to the Theory of Games*. vol. 2. Princeton, NJ, USA: Princeton Univ. Press, 1953.

[165] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.

[166] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowl. Inf. Syst.*, vol. 41, no. 3, pp. 647–665, Dec. 2014.

[167] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

[168] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[169] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[170] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K. R. Müller, "Layer-wise relevance propagation: An overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 193–209.

[171] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[172] D. Liu, P. Fengjiao, and R. Picard, "DeepFaceLIFT: Interpretable personalized models for automatic estimation of self-reported pain," in *Proc. IJCAI Workshop Artif. Intell. Affect. Comput.*, 2017, pp. 1–16.

[173] P. Prajod, T. Huber, and E. Andrá, "Using explainable AI to identify differences between clinical and experimental pain detection models based on facial expressions," in *Proc. Int. Conf. Multimedia Model.* Cham, Switzerland: Springer, 2022, pp. 311–322.

[174] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 542–552, Jul. 2020.

[175] S. Malik, P. Kumar, and B. Raman, "Towards interpretable facial emotion recognition," in *Proc. 12th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2021, pp. 1–9.

[176] H. Zhu, C. Yu, and A. Cangelosi, "Explainable emotion recognition for trustworthy human–robot interaction," in *Proc. Workshop Context-Awareness Hum.-Robot Interact. Approaches Challenges ACM/IEEE HRI*, Sapporo, Japan, Mar. 2022.

[177] T. Hwu, M. Levy, S. Skorheim, and D. Huber, "Matching representations of explainable artificial intelligence and eye gaze for human–machine interaction," 2021, *arXiv:2102.00179*.

[178] A. Heimerl, K. Weitz, T. Baur, and E. Andre, "Unraveling ML models of emotion with NOVA: Multi-level explainable AI for non-experts," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1155–1167, Jul. 2022.

[179] W. S. Liew, C. K. Loo, and S. Wermter, "Emotion recognition using explainable genetically optimized fuzzy ART ensembles," *IEEE Access*, vol. 9, pp. 61513–61531, 2021.

[180] K. R. Chowdhury, A. Sil, and S. R. Shukla, "Explaining a black-box sentiment analysis model with local interpretable model diagnostics (don't short) explanation (LIME)," in *Proc. Int. Conf. Adv. Comput. Data Sci.* Cham, Switzerland: Springer, 2021, pp. 90–101.

[181] I. Tenney et al., "The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models," 2020, *arXiv:2008.05122*.

[182] Y. Lyu, P. Pu Liang, Z. Deng, R. Salakhutdinov, and L.-P. Morency, "DIME: Fine-grained interpretations of multimodal models via disentangled local explanations," 2022, *arXiv:2203.02013*.

[183] P. Pu Liang et al., "MultiViz: Towards visualizing and understanding multimodal models," 2022, *arXiv:2207.00056*.

[184] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[185] S. Alisamir and F. Ringeval, "On the evolution of speech representations for affective computing: A brief history and critical overview," *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 12–21, Nov. 2021.

[186] S. Das, N. Nadine Lønfeldt, A. Katrine Pagsberg, and L. H. Clemmensen, "Towards interpretable and transferable speech emotion recognition: Latent representation based analysis of features, methods and corpora," 2021, *arXiv:2105.02055*.

[187] S. Alghowinem, T. Gedeon, R. Goecke, J. F. Cohn, and G. Parker, "Interpretation of depression detection models via feature selection methods," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 133–152, Jan. 2020.

[188] Y. Nishimura, T. Hossain, A. Sano, S. Isomura, Y. Arakawa, and S. Inoue, "Toward the analysis of office workers' mental indicators based on wearable, work activity, and weather data," in *Sensor- and Video-Based Activity and Behavior Computing*. Singapore: Springer, 2022, pp. 1–26.

[189] R. Lewis, Y. Liu, M. Groh, and R. Picard, "Shaping habit formation insights with Shapley values: Towards an explainable AI-system for self-understanding and health behavior change," in *Proc. Realizing AI Healthcare, Challenges Appearing Wild CHI*, 2021.

[190] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi, "Bottom-up and top-down: Predicting personality with psycholinguistic and language model features," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2020, pp. 1184–1189.

[191] X. Wang, J. He, Z. Jin, M. Yang, Y. Wang, and H. Qu, "M2Lens: Visualizing and explaining multimodal models for sentiment analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 802–812, Jan. 2022.

[192] B. Schölkopf et al., "Towards causal representation learning," 2021, *arXiv:2102.11107*.

[193] D. Schiff, B. Rakova, A. Ayesh, A. Fanti, and M. Lennon, "Principles to practices for responsible AI: Closing the gap," 2020, *arXiv:2006.04707*.

[194] M. Chromik, M. Eiband, S. T. Völkel, and D. Buschek, "Dark patterns of explainability, transparency, and user control for intelligent systems," in *Proc. IUI Workshops*, vol. 2327, 2019, pp. 1–6.

[195] U. Ehsan and M. O. Riedl, "Explainability pitfalls: Beyond dark patterns in explainable AI," 2021, *arXiv:2109.12480*.