machine learning & knowledge extraction

MDPI

Review

Not in My Face: Challenges and Ethical Considerations in Automatic Face Emotion Recognition Technology

Martina Mattioli [1,2,*,†,‡] **and Federico Cabitza** [3,4,‡]

[1] Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University, 30172 Venice, Italy
[2] Department of Control and Computer Engineering, Polytechnic University of Turin, 10138 Turin, Italy
[3] Department of Informatics, Systems and Communication, University of Milano-Bicocca, 20126 Milan, Italy; federico.cabitza@unimib.it
[4] IRCCS Ospedale Galeazzi-Sant'Ambrogio, 20157 Milan, Italy
[*] Correspondence: martina.mattioli@unive.it
[†] Current address: Via Torino 155, 30172 Venice, Italy.
[‡] These authors contributed equally to this work.

Abstract: Automatic Face Emotion Recognition (FER) technologies have become widespread in various applications, including surveillance, human–computer interaction, and health care. However, these systems are built on the basis of controversial psychological models that claim facial expressions are universally linked to specific emotions—a concept often referred to as the "universality hypothesis". Recent research highlights significant variability in how emotions are expressed and perceived across different cultures and contexts. This paper identifies a gap in evaluating the reliability and ethical implications of these systems, given their potential biases and privacy concerns. Here, we report a comprehensive review of the current debates surrounding FER, with a focus on cultural and social biases, the ethical implications of their application, and their technical reliability. Moreover, we propose a classification that organizes these perspectives into a three-part taxonomy. Key findings show that FER systems are built with limited datasets with potential annotation biases, in addition to lacking cultural context and exhibiting significant unreliability, with misclassification rates influenced by race and background. In some cases, the systems' errors lead to significant ethical concerns, particularly in sensitive settings such as law enforcement and surveillance. This study calls for more rigorous evaluation frameworks and regulatory oversight, ensuring that the deployment of FER systems does not infringe on individual rights or perpetuate biases.

Keywords: FER; emotion recognition; reliability; ethics

check for updates

Citation: Mattioli, M.; Cabitza, F. Not in My Face: Challenges and Ethical Considerations in Automatic Face Emotion Recognition Technology. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2201–2231. https://doi.org/10.3390/make6040109

Academic Editor: Abdulhamit Subasi

Received: 27 August 2024
Revised: 13 September 2024
Accepted: 26 September 2024
Published: 30 September 2024

1. Introduction

Human beings can universally recognize emotions conveyed through facial expressions. This bold—although not entirely undisputed—statement has been supported by many specialist contributions to the scientific literature on emotions, becoming widely known as the "universality hypothesis [1]". Despite being implicitly rooted in Western culture, mentioned in introductory psychological books, and assumed in movies and social institutions [1–4], evidence has highlighted variations in how human beings express emotions, as influenced by various factors [1,5–8]. For instance, cultural norms [1,5,9] and context [10] significantly impact how emotions are displayed and perceived. This diversity is attributed to culturally different "display rules", intensity ratings, and emotion labeling [9]. Furthermore, the situational context in which a facial expression occurs impacts how emotions are perceived [10]. Since emotions can be expressed in multiple ways, inferring how someone feels based solely on a limited set of facial expressions becomes problematic [5,8]. Thus, many scholars challenge this simplistic view of emotions and argue that emotions are better understood as flexible patterns shaped by cultural and social factors rather than as specific universal facial movements [5,7,10–13].

Nevertheless, the proponents of the "universality hypothesis" argue that a set of basic—or primary—emotions can be identified across the full range of possible emotions, forming the foundation for more complex emotional experiences, similar to how primary colors blend to create the full spectrum of colors. This "classical view" is also reinforced by the renowned (although not uncontroversial) experiments conducted by Paul Ekman and Wallace V. Friesen in different countries and cultures unexposed to Western influence, which assert that a specific emotion can be linked to precise facial expressions by virtue of its universal nature. This implies the biological basis of facial expressions and their distinctiveness from culturally specific behaviors [14].

Despite the widespread acceptance of these claims across different domains [1,2,7], it is essential to delve into the question of how capable humans are of accurately inferring emotions from facial expressions. The relevance of this issue increases when universal assumptions underpin the development of FER systems, which autonomously infer private and subjective states from facial expressions. It should be mentioned that the acronym FER is employed differently in the pertinent literature to denote different terminology. For instance, FER often refers to face expression recognition. However, in this paper, we employ the acronym with reference to face emotion recognition. Such technologies are part of Affective Computing (AC), a subfield of Artificial Intelligence (AI) focused on detecting emotions from various data sources, such as speech signals, facial expressions, and textual inputs [15]. FER specifically involves a range of methods, primarily based on Machine Learning (ML) techniques, trained on large datasets of labeled images. In its ground-truthing process, a variable number of human raters is asked to annotate face pictures using one or more labels according to an emotion model that typically employs either categorical or ordinal values [16].

However, there is considerable debate within the scientific community regarding the feasibility of this task by virtue of findings backing the conclusion that emotions cannot be measured or treated as "entities" and that agreement in assigning emotion labels can be inconsistent [1,6,7,17,18], resulting in the poor reliability of such annotations [16].

Besides (and beyond) the technical feasibility of associating facial expressions with specific (and true) emotions, some scholars have raised ethical concerns about the nature of this approach. Psychological research has demonstrated that individuals tend to display a racial bias when attributing emotions to others, particularly in cases that involve negative emotions [19–22]. Specifically, this bias consists of the misperception of facial expressions due to racial prejudice, such as a tendency to associate ambiguous aggressive expressions more strongly with African Americans [19]. This correlation becomes notably problematic and harmful when FER systems are deployed, for example, in surveillance settings to forecast or oversee potentially offensive behaviors [23–25]. Racial stereotypes can lead to significant and troubling ethical implications, as they may be perpetuated within the system through annotations, resulting in biased AI [26]. By way of illustration, an emotional analysis conducted by Face++ and Microsoft AI on images of professional basketball players revealed a tendency to assign negative emotions to Black players as compared to their White counterparts, showing that facial recognition software exhibits variations in emotion interpretation influenced by the individual's race [26]. The employment of this technology can not only result in racial biases or stereotypes but can also infringe on privacy and human rights. For instance, Zoom's emotion recognition technology has drawn criticism from advocacy groups. In an open letter to Zoom's CEO, these groups expressed concerns that AI emotion systems could monitor users' facial expressions and emotional responses without their consent, leading to potential misuse of personal data [27].

Aim and Scope of the Study

Given the broad and varied nature of discussions surrounding FER applications, we believe it is essential to elucidate the various positions that have arisen in the debate. In recent years, there has been an increasing interest in interdisciplinary research about emotion recognition technologies and their implications [28,29], resulting in a substantial body of literature that reflects diverse perspectives on FER [16]. Our article aims to summarize and review different aspects of FER concerns based on a review of 96 papers. To address the multifaceted nature of the debate, we developed a conceptual taxonomy that organizes the key perspectives identified in the literature. These concerns span various topics, ranging from strong skepticism about the possibility of accurately inferring emotions from proxy data like facial expressions [2,6,30] to more moderate positions about the ethical implications of deploying these systems in sensitive contexts [28] and questions surrounding the reliability of FER [16]. Consequently, this article provides a comprehensive review of the field, exploring the wide range of arguments and perspectives, which we categorize into three main areas of critique, namely the questionable psychological foundations of FER, its negative ethical impact, and the reliability of its ground truth. In doing so, we show how different levels of analysis impact different aspects of FER technology.

The purpose of this paper is to serve as a "single-point resource" for understanding the recent controversies, debates, technical challenges, and ethical considerations surrounding FER technologies, thereby offering a comprehensive examination of the divergent perspectives and arguments within the field, as shown in Figure 1. By addressing these key issues, we aim to enhance the understanding of the nuanced concerns surrounding various aspects of FER. In summary, the key objectives of this paper are to

- Examine the psychological and philosophical debate on the scientific feasibility of accurately inferring human emotions from facial expressions;
- Present an overview of how FER technology works, focusing on critical applications, failures, and misclassification errors;
- Review FER datasets by classifying them as those with simulated vs. genuine expressions and those based on categorical vs. ordinal models, highlighting concerns regarding bias and reliability;
- Critically review the literature surrounding FER concerns, focusing on their psychological foundations, ethical consequences, and reliability issues;
- Propose a conceptual taxonomy that categorizes the major critiques and discussions about FER technologies mentioned above, facilitating a structured and comprehensive understanding of the field.

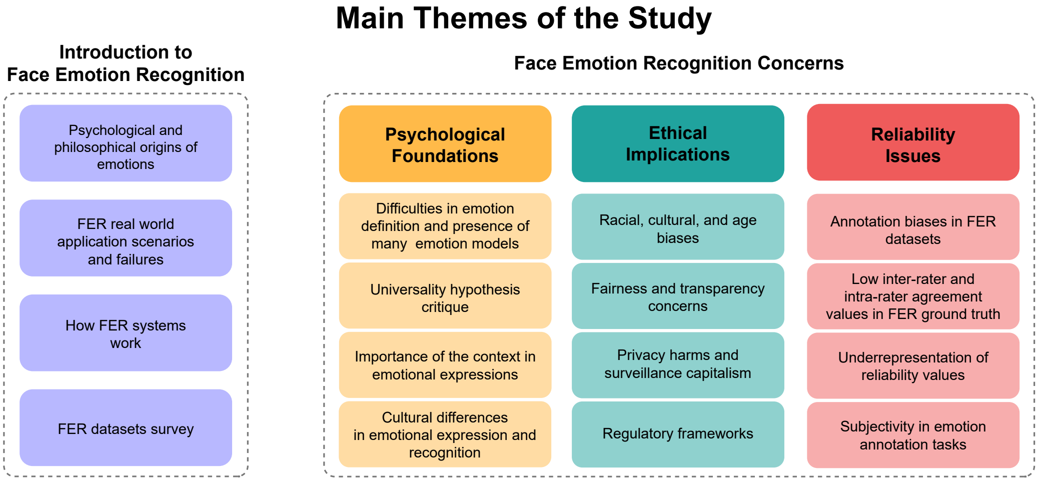

Main Themes of the Study
Introduction to Face Emotion Recognition
Psychological and philosophical origins of emotions
FER real world application scenarios and failures
How FER systems work
FER datasets survey
Face Emotion Recognition Concerns
Psychological Foundations
Difficulties in emotion definition and presence of many emotion models
Universality hypothesis critique
Importance of the context in emotional expressions
Cultural differences in emotional expression and recognition
Ethical Implications
Racial, cultural, and age biases
Fairness and transparency concerns
Privacy harms and surveillance capitalism
Regulatory frameworks
Reliability Issues
Annotation biases in FER datasets
Low inter-rater and intra-rater agreement values in FER ground truth
Underrepresentation of reliability values
Subjectivity in emotion annotation tasks

Figure 1. Overview of the study's key findings. The left part presents the introductory topics of FER discussed herein, while the right part illustrates our classification of primary areas of concern with respect to FER .

To achieve this, the paper is organized as follows. Section 2 describes the methodology of the review. In Section 3, because of the peculiar and troublesome historical evolution of emotions [31], we trace their origins and highlight how different historical periods and psychological models have shaped their study. Subsequently, in Section 4, we provide a high-level overview of FER and its applications. Finally, in Section 5, we organize and discuss the debate around three predominant perspectives identified in the literature. Specifically, we present the contribution of authors who deny the scientific feasibility of inferring human emotions from facial expressions. Then, we underline the critique of the employment of FER applications in real-life scenarios as potentially harmful to human rights. Finally, we address the contentions according to which FER datasets are often biased and unreliable.

2. Methodology

To conduct a comprehensive review of FER technologies, we implemented a systematic approach to identify, analyze, and categorize relevant academic literature. A curated set of FER-related keywords was used for database searches of the Web of Science and Google Scholar aimed at retrieving foundational research on the core mechanisms of FER. In addition, we combined these terms with keywords like "critique", "issues", "criticism", "concerns", "harms", and "ethics" to locate studies that focus on critical evaluations of FER. This search formed the initial basis for our literature collection, which was further expanded through a citation analysis of the references in the included studies to uncover additional pertinent articles.

Furthermore, we consulted classical psychological manuals to source psychological research on emotions, offering guidance on FER's theoretical underpinnings. Applications and real-world cases of FER were drawn from both daily news reports and scientific databases, including the Google search engine, Google Scholar, and Web of Science.

Following the literature collection, we developed a conceptual taxonomy to systematically organize critical perspectives on FER. This taxonomy emerged through the thematic grouping of articles based on recurring critiques and concerns identified in the literature. The following three primary categories were defined:

- Psychological Foundations: Articles that critique or challenge the universality hypothesis and explore how cultural and social factors influence emotion recognition.
- Ethical Implications: Studies addressing biases and ethical concerns in the deployment of FER systems, with emphasis on issues such as racial bias, privacy, and human rights.
- Reliability Issues: Research evaluating the reliability and accuracy of FER datasets, particularly in terms of inter-rater and intra-rater reliability in human-annotated data.

For inclusion, we focused exclusively on articles centered on facial expression analysis, despite the existence of other, often more successful methods for emotion detection through the use of biosignals, such as skin conductance [32] or EEG (electroencephalogram) [33], thereby excluding studies that addressed broader aspects of emotion recognition. In total, we reviewed 96 articles, 16 datasets, 6 daily news articles, and 6 legal sources.

3. Finding the Origins of Emotions

In 1884, William James posed the following question: "What Is an Emotion"? [34]. One hundred and forty years later, this interrogative remains relevant, while, with respect to the answer, there is still little scientific consensus [6,31,35]. The understanding of emotions has undergone a complex evolution across the realms of philosophy and cognitive sciences, involving a multitude of thinkers, all seeking a definitive definition. The numerous difficulties in this task are determined by various factors, including the intrinsic complexity of the study of emotions (being private and subjective states) [35] and their negative historical connotations [31]. Indeed, they have been traditionally described as violent forces antithetical to rationality and, consequently, regarded as a status to be avoided [31]. Moreover, the development of this word shares its roots with satellite terminology, such as "passion" and "affections", which were shaped by Christian philosophical ideas that saw reason and emotions as opposing entities [31]. For instance, Saint Augustine and Thomas Aquinas viewed "passions" as potent appetites capable of conflicting with reason and leading individuals toward sin. In contrast, “affections” referred to those mental states guided by rationality and were deemed to bring individuals closer to God [31].

Therefore, the study of emotions can be described as one of the most confusing chapters of psychological research [36], having produced more than ninety definitions over the centuries, making current investigations chaotic and complicated [37]. Confirming the complexity of this term, the literature does not provide a clear and unique definition but multiple operational delineations aimed at coherently presenting different aspects of emotions [35], which we summarize in the following subsections.

3.1. The Main Psychological Models of Emotions

The search for an answer to the question posed by James [34] has occupied numerous scholars from diverse psychological traditions, resulting in a very large number of model proposals [35]. In the early behaviorist theories, emotions were considered straightforward products of behavior, as behavior itself was regarded as the sole reliable source of information. Following this perspective, science was not able to link basic patterns of bodily responses to specific internal states, since they are non-diagnostic [38]. Hence, emotions were examined as mere reactions, such as “sadness” corresponding to crying and “happiness” corresponding to smiling [2]. However, this type of theory neglects a fundamental fact, i.e., that humans elaborate emotions in their brains and cognition through underlying processes [2,35]. As a reaction against this view, cognitive scientists prioritized the exploration of the organization of the mind to understand the causes and effects of these inner processes. They considered emotions to be “value judgments”, meaning they are not merely physical responses but involved an appraisal activity. Cognition evaluates real, imaginary, or abstract events, and an emotional response follows this evaluation process [39]. In contrast, psychoanalytic models were founded upon the identification of emotions for the purpose of therapeutic practice. In many instances, these models are indirect in the sense that they address emotions only secondarily within a larger theoretical framework [35]. For instance, Sandor Radó [40] posited that emotions can be discerned through a multitude of indicators, including behavior during therapy, free associations, and dreams. Moreover, he argued that it is often possible to detect the presence of emotions in subjects who do not believe they are experiencing them. However, it is mainly because of the contribution of the “cognitive revolution” [2] that motivational theories (i.e., Basic Emotion Theories (BETs)) arose, carrying forward their strong statements about the universality of emotions. This perspective is rooted in evolutionary theory, positing that an animal’s emotional expression traces back to behaviors crucial for advancing the survival of the species [35]. Also known as the “classical view” [5] due to its paramount significance in emotion research, it hypothesizes that a small set of basic emotions can be identified and tracked through specific elements of facial or bodily behavior, as well as general proxy data, which are cross-cultural [14]. The implications and influence of BETs have transcended psychological research, extending into diverse domains such as AI and law and serving as conceptual foundations for the design of ML algorithms [5,28]. Indeed, the employment of basic emotions lends itself to facile application due to the straightforwardness inherent in implementing this particular emotional model within artificial emotion recognition systems [28].

3.2. Measures and Labeling of Emotions: What Is the Role of Facial Expressions?

For the purposes of the present discussion, we temporarily and completely accept universal assumptions and that emotions can be reliably inferred and measured from facial expressions or other proxy data. More simply, we assume that private states stand in a one-to-one relationship with expressions in every human individual. Without casting doubt on this hypothesis, the question of how emotions can be measured becomes essential. Their assessment and measurement are contingent upon the theories and models of reference. For instance, motivational theories, which are grounded in the assumptions of universal expressions of emotions, tend to focus on observing facial expressions and other physiological changes [35]. On the other hand, psychoanalytic models mainly refer to self-description, as they are predominantly employed in clinical practice [35]. However, in general, in the psychological literature, scholars typically identify four different methods to measure emotional feelings within a human subject, namely self-descriptions of subjective experiences, evaluation of behavior, assessment of the product of behavior, and recording of physiological changes [35]. In the latter category, various methods can be mentioned, including EEG, metabolic rate, skin sweating, and respiration rate [35].

In particular, the study of facial expressions has garnered substantial interest from psychologists and researchers across various disciplines [35]. A vast number of theories have been developed to elucidate the relationship between expressions and emotions. These can be broadly categorized into the following two main groups: peripheral and central theories. Peripheral theories posit that the sole act of contracting facial muscles generates feedback, which subsequently elicits emotions. In contrast, central theories argue that facial expressions mirror the individual's emotional state [35]. Studies such as the research conducted by Ekman and Friesen [14] are key examples of the latter category. These authors sought to demonstrate the presence of cultural invariants by analyzing expressions of emotions. By doing so, they focused mainly on the face and used images that were presumed to be illustrative of inner states, asking participants to pair them with labels referring to a precise emotion (each label describing a basic emotion). Such experimental settings vividly exemplify the underlying principles of central approaches, which posit a direct correlation between expression and emotion and that have been replicated analogously by experiments conducted to construct datasets for the training of FER systems [41].

Additionally, one can distinguish between discrete categorical and continuous approaches. The categorical approach involves assigning specific labels to fundamental emotions, whereas the continuous approach utilizes multi-dimensional and ordinal measures. Silvan Tomkins [42], for instance, proposed a list of eight primary affects divided between positive (interest–excitement, enjoyment–joy), resetting (surprise–startle), and negative affects (distress–anguish, fear–terror, shame–humiliation, contempt–disgust, and anger–rage) described as innate structured responses that occur through different physiological variations in which the face is seen as the most important site for the expression of subjective states. Robert Plutchik [35] argued that it is possible to identify eight emotional dimensions at the phylogenetic level that are of adaptive significance for the species' survival, namely anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. Ekman and Friesen [14] identified six basic emotions (anger, surprise, disgust, enjoyment, fear, and sadness) that are distinctive universal signals and are marked by unmistakable facial expressions.

Built upon basic emotions, secondary emotions are a combination of basic emotions, and their composition depends on the primary emotions that are involved, which can be mixed in different ways. To maintain the color analogy, according to this view, a finite number of basic emotions exist, which should be fixed and known a priori, and, when mixed correctly, should give rise to a particular secondary emotion [2]. Nonetheless, in the literature, there is not a clear consensus about which emotions should serve as basic emotions [35]. Indeed, the delineation of both the quantity and the specific emotions encompassed within this category has sparked intense debate [35]. Furthermore, the criteria for determining primary or secondary emotions are not well defined, leading to considerable variation in attempts to list basic emotions within motivational theories [35].

Due to the challenges of representing a large number of emotions within a framework consisting of only a small number of basic emotions, multi-dimensional and continuous approaches have been proposed, providing more sophisticated methodologies for capturing nuances in the interplay of diverse emotional states and their respective intensities [43]. Wilhelm Wundt [44] was one of the pioneering psychologists to propose that subjective experiences comprise at least two properties, namely valence (from unpleasant to pleasant) and arousal (from calm to active). For instance, anxiety and depression are both associated with negative emotions, but they differ in their level of arousal. High levels of arousal characterize anxiety, while low levels describe depression. In the literature, it is possible to identify two- and three-dimensional frameworks to depict emotions. In the former case, emotions are represented through two dimensions (e.g., valence and arousal), while in the latter case, they are represented through the use of three dimensions (e.g., valence, arousal, and dominance) [35]. The circumplex model developed by James Russell [45] is one of the most popular two-dimensional models. It includes eight variables represented in a two-dimensional circle. The horizontal dimension (x) represents pleasure/dislike, and the vertical axis (y) represents arousal/calm. The model also includes four other variables (distress, excitement, contentment, and depression), which are located in the quadrants produced by the intersection between the axes, as shown in Figure 2. The Valence–Arousal–Dominance (VAD) model [46] is commonly used to identify emotions in a three-dimensional space. It measures valence on a scale ranging from negative emotions (not happy) to positive emotions (happy). Arousal is measured on a scale ranging from calm emotions to energetic emotions. Dominance is measured on a scale from "without control" to "under control". In addition to VAD, there are other three-dimensional models, such as PAD (Pleasure–Arousal–Dominance) [47], in which the first dimension measures pleasantness rather than valence. Finally, the Facial Action Coding System (FACS) categorizes facial expressions based on the activation of specific facial muscles known as Action Units (AUs). The combination of different AUs results in a different emotions [48].

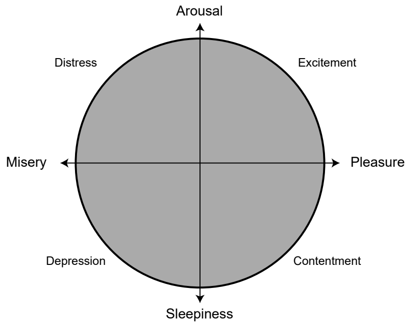


Figure 2. The circumplex model developed by Russell. It includes eight variables represented in a two-dimensional circle. The horizontal dimension (x) represents pleasure/dislike, and the vertical axis (y) represents arousal/calmness [45].

Regardless of whether these models are categorical or ordinal and whether they track facial expressions or other physiological changes, they all share the underlying assumption that emotions are quantifiable entities [6,17,18]. Events are posited to trigger specific emotions, which, in turn, produce a set of assessable behaviors. These behaviors can then be measured and labeled, and the occurrence of an expression serves as clear evidence of the presence of an emotion. Indeed, the belief fundamental to the idea of inferring and measuring emotions through proxy data and deriving the "universality hypothesis" is the existence of "fingerprints" that should provide the objective specification of the correct identification and measurement of emotions. However, some authors have argued that the empirical evidence produced to validate such models suggests that it is complicated, if not impossible, to find an objective measure to evaluate the experience of emotion [6,17,18].

4. An Overview of Face Emotion Recognition

Emotion recognition systems belong to AC technology, an interdisciplinary field encompassing ML and cognitive science that seeks to develop systems that can perceive, reproduce, simulate, and understand human emotions [15]. Specifically, the recent final draft of the AI Act defines

> *"The notion of 'emotion recognition system' [...] as an AI system for the purpose of identifying or inferring emotions or intentions of natural persons on the basis of their*

> *biometric data. The notion refers to emotions or intentions such as happiness, sadness, anger, surprise, disgust, embarrassment, excitement, shame, contempt, satisfaction and amusement* [49]."

In the realm of AI systems, practitioners often rely on various types of proxy data to analyze emotional experiences. These generally encompass facial expressions, physiological data, skin conductivity, blood flow, body movement, audio data, and EEG [50]. Specifically, FER involves identifying and interpreting facial expressions to infer and categorize emotional states or intentions, thereby deciphering one of the most important media through which humans communicate intentional emotions, namely the face [15,35]. Precisely, automatic facial emotion recognition relies on the premise of asserting the existence of specific somatic patterns accompanying each basic emotion, attributing to them a distinct and identifiable array of muscle movements [28]. Despite a substantial body of psychological literature proposing the absence of a straightforward relationship between emotion and expression [2,6,17,18], systems for facial emotion recognition primarily depend on these delineated affective schemes that are discernible within non-verbal communicative language [28]. However, before delving into an in-depth discussion of FER concerns throughout this paper, this section provides a general and concise overview of FER to establish the foundation necessary for the subsequent analysis of the various positions regarding its concerns. Specifically, we provide examples of FER in real-world scenarios and a high-level overview of FER techniques. Additionally, we present some frequently used datasets.

4.1. FER Applications: Examples from the Real World

FER is employed across a broad spectrum of contexts, encompassing various fields and disciplines. However, the perspectives concerning its applications exhibit significant divergence [16], which can be attributed to various factors, such as the lack of reliability [16,51], questionable theoretical foundations [5,6,52], and the ethical considerations associated with its use [28,50], which we discuss in Section 5. Nevertheless, it can be beneficial to have a comprehensive understanding and conduct a critical evaluation of different perspectives on FER application examples for a more thorough analysis.

A segment of scholars underlines FER's potential benefits and good performance in different areas. For example, we mention car driving safety [53], support for individuals with communication deficits [54], airport and public space surveillance [55], support for the selection of human resources [56], human–computer interaction [57], and medicine [58]. For instance, Kalpana M. Chowdary et al. [57] emphasized the benefits of AC in human–computer interaction in applications such as online teaching, virtual sales assistants, internet banking, medicine, and security. The authors reported high accuracy in facial emotion recognition using pre-trained Convolutional Neural Networks (CNNs). Specifically, they used VGG19 [59], ResNet50 [60], Inception V3 [61], and MobileNet [62] models trained on the ImageNet database [63] and tested on the CK+ database [64]. The achieved accuracies were 96% for VGG19, 97.7% for ResNet50, 98.5% for Inception V3, and 94.2% for MobileNet.

Additionally, among the applications mentioned above, it is worth deepening the point of view of those researchers who argue that artificial emotion recognition would be valuable in supporting therapy and improving communication [54,65,66], such as in Autism Spectrum Disorders (ASDs). To illustrate this perspective, Rodolfo Pavez et al. [65] proposed an "intelligent mirror" to help children with ASDs recognize five basic emotions, as well as analysis and comparison of images captured by the system camera. The authors conducted a specific evaluation of the VGG16 [59] and ResNet50 architectures by resizing images to 200 × 200 pixels and utilizing Stochastic Gradient Descent (SGD) for optimization. The models were trained on both the FER2013 [67] and CK+ [64] datasets. After an evaluation with professionals, Pavez et al. [65] concluded that the prototype of this technology could assist ASD children and autism specialists. Although these and similar applications have been widely implemented, some scholars have highlighted the limitations of FER and encouraged the combination of facial expressions with additional indicators such as EEG to avoid the disadvantages of the employment of facial expression pictures, such as the potential for deception or poor lighting conditions [33]. However, others argue that there remains a lack of alignment between physiological data—used as a proxy—and an individual's private emotional experience [68,69].

Moreover, specific application contexts, such as public space surveillance or scenarios that are intrusive to an individual's privacy, are deemed more sensitive by some, warranting a more cautious deployment of these technologies and highlighting potential harms and instances of failure observed in its applications [50,69,70]. For instance, Lena Podoletz [70] emphasized the importance of fully comprehending the implications, challenges, and limitations associated with the deployment of emotional AI technologies in policing. Specifically, she argued that crime predictions based on probability cannot make definitive assessments of future events. The grounding ideas of the notion of predicting crimes based on behavioral patterns are derived from the conviction of the existence of signs that precede the commission of a criminal act. However, the author maintains that, currently, it remains uncertain whether one can conclusively determine people's emotional states solely based on their behavior. For instance, Flavia Spiroiu [71] investigated how false beliefs might shape the perception of non-verbal behavior of individuals. Her study revealed that participants' reports on eye movements were influenced by the information they received about a suspect's guilt and the alleged link between eye movements and deceptive behavior. Specifically, those informed that "liars look left" reported significantly more leftward eye movements for suspects labeled as "deceptive", even though the actual eye movements were mainly directed to the right. Additionally, the notion of predicting crimes according to behavioral patterns is tied to the outdated belief that identity traits, like criminality, can be inferred from physical appearance. This mirrors physiognomy, a discriminatory pseudoscience practiced until the late nineteenth century that associates specific facial features with personality [3].

Furthermore, employing such surveillance methods infringes on individuals' rights to privacy regarding their location, thoughts, and emotions [70]. Some examples of applications that have had a negative impact on privacy and other personal rights come from real-world applications of such technologies. Of particular gravity and regrettably well-known to public opinion is the case of the AI-integrated surveillance system that is widespread in the Xinjiang region [72], in which the minority of the population are Uyghurs, which

> "*Has effectively become a 'frontline laboratory' for data-driven surveillance*" [73].

Indeed, Uyghurs belong to a Muslim Turkic-language-speaking minority that lives in the Xinjiang region and, for a long time, have been subjected to Chinese repression. Specifically, facial recognition has served as a means of oppression, entailing constant surveillance and control of citizens under the guise of ensuring safety, maintaining order, and advancing societal development [72,74]. Moreover, we mention the iBorderCtrl project, which trialed an AI-based lie detector at European borders to assess travelers' potential risks. Using facial recognition and micro-expression analysis, the system aimed to detect deception during border checks. However, the technology faced criticism due to concerns over its accuracy, bias, and infringement of privacy and individual rights. Critics argued that such tools can exacerbate discrimination, particularly against minority groups, and present significant ethical and human rights challenges, resulting in the withdrawal of this technology [75].

In general, the utilization of biometric data, including facial features, body measurements, and facial expressions, can result in significant misclassification errors. An unsettling example comes from a recent case in Brazil, where a facial recognition system misidentified a man in a stadium crowd in Sergipe, leading to his public escort by authorities. This incident highlights the potential for false positives in biometric systems and underscores the serious implications for individual rights and freedoms [74,76]. In particular, FER technology detects the most probable emotion among possible emotions [16]; nonetheless, these probabilities reflect degrees of certainty rather than absolute conclusions [70].

Moreover, this inherent uncertainty is especially pronounced in emotion recognition technologies due to various factors, such as the employment of proxy data, data processing into feature vectors, and the interpretation of the results [77]. The following section provides general information about the technologies underlying emotion recognition systems and the aforementioned applications.

4.2. How Does FER Work?

FER is not a unitary field of study and comprises a multitude of approaches and technologies. For instance, Felipe Z. Canal et al. [78] and Byoung Chul Ko [79] distinguished between conventional and Deep Learning (DL)-based approaches. Specifically, the former encompasses methods that rely on hand-crafted features, while in the latter, the features are learned from the data. Moreover, the process of identifying facial emotions involves several distinct stages, namely data acquisition, pre-processing, face detection, feature extraction, and classification algorithms [78,80,81]. This process is illustrated in Figure 3.

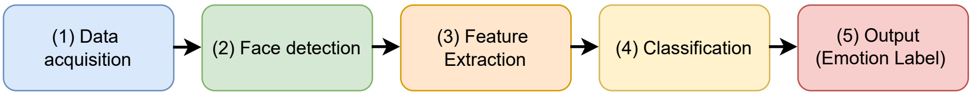


Figure 3. Graphic representation of the computational process of FER.

Generally, the initial stage of facial emotion recognition is pre-processing, which is carried out with the objective of enhancing the quality of input data. The tasks incorporated into this stage typically include noise reduction, image resizing, and normalization procedures with the objective of standardizing the images [78]. Following the pre-processing stage, face detection algorithms, such as the Viola-Jones algorithm [82], the Haar Cascade Classifier [83], and Adaboost Contour Points [84], are employed to identify faces within the given input images or video frames. After a face is identified, its defining characteristics (e.g., eyes, eyebrows, nose, mouth, and chin) are extracted. A few classical algorithms for this task are Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Active Shape Models (ASMs) [78]. Subsequently, the extracted features are classified into distinct emotional states. This is typically accomplished through the utilization of traditional ML techniques, such as Support Vector Machine (SVM) [85], Random Forest (RF) [86], and Dynamic Bayesian Networks (DBNs) [87]. More recently, DL architectures such as CNNs and Recurrent Neural Networks (RNNs) have gained prominence for the tasks of both feature extraction and classification [78,79]. The growing use of DL technologies can be attributed to several factors, including, for example, the transition from laboratory-controlled to in-the-wild settings, the vast abundance of data, and the enhanced accessibility of computational resources [79,88,89]. Indeed, DL has achieved state-of-the-art results in FER [60,88–90]. In traditional ML methods, feature extraction and classification are often treated as separate steps. Handcrafted features are extracted first; then, a classifier is applied. In contrast, DL-based approaches integrate the entire process, including feature extraction and classification, into a single model [89]. In the specific context of FER, for instance, Deep Neural Networks (DNNs) are trained to map input facial images directly to output predictions of facial expressions [88–90]. This is accomplished through the use of multiple layers of interconnected neurons, which are capable of automatically learning hierarchical representations of the input data. A loss layer is added at the end of the network, which calculates the error between the predicted outputs and the ground-truth labels. Subsequently, the error is propagated backward through the network, enabling the model's parameters to be adjusted in order to minimize the error [89]. Hence, the classifier learns the intricate relationships between the extracted features and their corresponding emotional labels during training. Subsequently, the learned relationships are generalized to predict emotions for test samples based on features extracted from previously unseen facial expressions.

Following the stages already mentioned, when evaluating FER systems, a variety of metrics is employed to assess their performance and provide a benchmark for comparison. Frequently used metrics include accuracy, recall, precision, and average processing time [79,80]. However, besides algorithm performance, this technology relies on labeled datasets commonly consisting of human-annotated images based on a distinct emotion model and measure, such as a categorical or dimensional model. Indeed, various authors also refer to the dataset's quality when evaluating FER [16,80]. Since FER depends on ML algorithms, its accuracy is inherently tied to the reliability of its ground truth [16,91,92]. Because these datasets have a vast number of images, the labeling task is typically undertaken by multiple raters who are given different expression images "in the wild" (i.e., pictures downloaded from various Internet sources) or images depicting actors simulating an expression [41]. Nonetheless, human annotation might hold some subjectivity or biases, rendering the task of assigning labels to emotions uncertain [77]. To complete this descriptive overview, in the following section, we present an outline of the main datasets available in the literature.

The recognition of emotion through data-driven learning is contingent upon the availability of appropriate datasets. In the literature, a wide variety of databases is employed as ground truth for the training of FER algorithms [16]. Indeed, datasets differ on the basis of several factors, such as the number of images, the type and number of annotators, the source of the collected images, and the emotion model employed [16,41]. There are apparent differences in how these datasets are built, which are explained by so-called "capture bias", which is the preference expressed during their construction [93]. For instance, early datasets often comprised images taken in controlled environments, such as laboratories, where facial expressions were elicited by stimuli or resulted from voluntary behavior [41]. Moreover, images can be shot with different types of cameras, while pictures "in the wild" contain a broader range of genders, ethnicities, and ages [41,93]. Additionally, datasets may present different types of bias or racial prejudices that can be reproduced within the systems themselves [28]. This section briefly introduces the principal datasets containing pictorial information published after 2010 that were identified in the literature and classified according to the type of images (simulated or fake expressions or genuine expressions) and the model employed to measure emotions (categorical model or ordinal model).

Categorical Lab Setting Datasets

This category encompasses datasets in which facial expression were simulated within a controlled environment and subsequently classified with a variable number of labels of basic emotions. Typically, the earliest datasets were acquired within a laboratory setting, comprising subjects exhibiting various facial expressions under controlled conditions. Consider, for example, the Tsinghua dataset [94], in which participants (only Chinese individuals) were asked to pose for a specific expression, with the most representative images for each category subsequently selected. This method produced a refined, high-quality dataset of intentionally posed facial expressions [41]. However, it is important to acknowledge that posed expressions may not fully align with the natural, spontaneous facial expressions observed in everyday contexts [2,41].

The DDCF [95] dataset presents the facial expressions of children. It features 80 Caucasian subjects aged 6 to 16 years, comprising 40 female and 40 male individuals. As anticipated, this dataset comprises facial expressions of posed individuals captured from five different angles. Each model was asked to pose in eight different facial expressions, as shown in Table 1. The dataset was validated by Dartmouth College students across seven different labels (neutral, content, sad, angry, afraid, disgusted, and surprised). Judges were asked to rank facial expressions in relation to the previous labels and the age of the subject present within the photograph. On average, expressions were correctly recognized 79.9% of the time, with a standard deviation of 22.7%. Cohen's coefficient [96] was calculated to be $\kappa = 0.780$.

Table 1. List of datasets that encompass images sourced from laboratory settings employing categorical values.

Dataset	Basic Emotion Labels
DDCF [95]	Neutral, Content, Sad, Angry, Afraid, Happy, Surprised, Disgusted
CAFE [97]	Sadness, Happiness, Surprise, Anger, Disgust, Fear, Neutral
NVIE [98]	Happiness, Disgust, Fear, Surprise, Sadness, Anger
TSINGHUA [94]	Neutral, Happiness, Anger, Disgust, Surprise, Fear, Content, Sadness
DEFSS [99]	Happy, Sad, Fearful, Angry, Neutral, None of the Above

CAFE [97] is a dataset comprising 1192 photographs of children's faces (aged 2–8 years), representing a diverse range of ethnicities, including Caucasian and non-Caucasian individuals. In fact, the dataset includes African American, Asian, Hispanic, and South Asian individuals. This database is based on the work conducted by Ekman and Friesen in 1971 [14], in which the authors employed six basic emotions. The photographs of children were taken within a controlled laboratory environment, with the photographer requesting the children to replicate their expressions. The dataset was validated by 100 adult subjects, with an accuracy rate of 66%.

The NVIE [98] dataset distinguishes itself from previously illustrated datasets by presenting a collection of images captured through infrared from a thermal cameras, aiming to mitigate concerns associated with varying lighting conditions. Encompassing both genuine and fake expressions, the dataset underwent validation by five judges for depicted emotions, excluding those portraying fake expressions. The authors reported Cohen's $\kappa = 0.65$ for inter-rater reliability within the dataset.

The Tsinghua [94] dataset comprises a collection of facial expressions of native Chinese subjects, a demographic group that is often under-represented in FER databases. In addition, facial expressions were gathered from subjects identified as young and old in controlled laboratory settings. Participants were instructed to exhibit facial behavior in accordance with the researchers' requests based on the basic emotions identified by Ekman and Friesen, as shown in Table 1. Figure 4 shows samples from the Tsinghua dataset. The dataset was validated by 60 raters with an inter-rater agreement corresponding to Cohen's $\kappa = 0.761$.

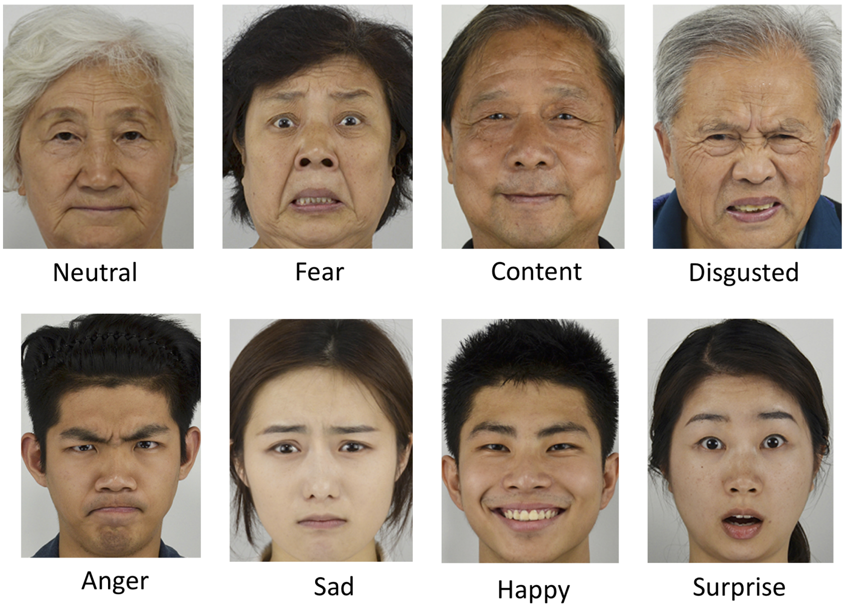


Figure 4. Samples of images and their related labels from the Tsinghua dataset [94].

4.3. The Construction of FER Datasets

DEFSS [99] is a dataset that contains images of faces belonging to individuals between the ages of 8 and 30 years. The dataset comprises 404 photographs, each labeled according to one of five basic emotions, as shown in Table 1. In order to represent the emotions, subjects were asked to manifest the expression they would feel in the case of a scenario described by the researchers. The entire set underwent a validation process involving 228 judges, who were asked to assess the expressions in relation to the five previously mentioned emotions.

4.3.1. Ordinal Lab Setting Datasets

This category encompasses datasets in which images were collected in controlled settings and the measurement of emotion was registered through ordinal values.

The DISFA [100] dataset, for instance, contains 4845 frames of 27 adults aged 18–50 years with a predominantly Caucasian ethnicity. These frames were manually annotated by a single person through AU, measuring intensities from 0 to 5. The subjects were required to view a video clip of approximately four minutes in duration, with the objective of evoking a specific emotional response within a controlled setting.

The Radboud Faces Database [101] comprises images of 49 Caucasian models, who were asked to display eight facial expressions, as detailed in Table 2. These expressions were identified as those most commonly recognized in experiments conducted by Ekman. A total of 120 photographs were taken per subject (with three different gaze directions), and during the photo sessions, the models were instructed by FACS experts. The images were subjected to validation by 276 students, who were asked to assign values to the dimensions of intensity, clarity, genuineness, and valence on a five-point scale. This process yielded an overall agreement of 82%, although there was greater disagreement and overlap in the case of expressions such as surprise and fear.

Table 2. List of datasets that encompass images sourced from laboratory settings employing ordinal values.

Dataset	Basic Emotion Labels
DISFA [100]	Action units with intensity from 0 to 5
Radboud Faces Dataset [101]	Neutral, Anger, Sadness, Fear, Disgust, Surprise, Happiness, and Contempt Dimensions: Clarity, Intensity, Genuineness, and Valence.

4.3.2. Categorical In-the-Wild Datasets

This dataset category comprises images capturing spontaneous expressions "in the wild", featuring individuals not prompted to pose with specific emotions within controlled laboratory settings. Labeling typically follows a categorical approach. These datasets predominantly consist of larger collections sourced from the Web [41].

The FER-2013 [67] dataset was built by employing the Google search engine to source facial images, pairing them with a set of 184 emotion-related words alongside additional keywords like gender, ethnicity, and age. This process yielded around 600 search results, with the first 1000 images returned by Google being utilized. A total of 35,887 images were collected and categorized into seven emotions, as detailed in Table 3. Figure 5 depicts samples from this dataset. Human labelers curated the dataset, excluding mislabeled images and refining bounding-box clipping when necessary.

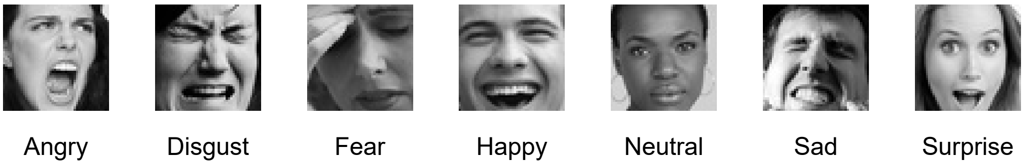


Figure 5. Samples of images and their related labels from the FER-2013 dataset [67].

The Google-FEC [102] dataset is structured to comprise triplets, wherein each sample consists of three images annotated with corresponding terms. These images are associated with at least one label listed in Table 3. The pictures were selected in such a way that all of the categories were represented to a greater or lesser extent. Each triplet underwent evaluation by five judges, resulting in a total of 40 evaluators. The following three distinct triplet types were delineated: one-class triplets, wherein all images share a single label; two-class triplets, wherein two images share a label; and three-class triplets, wherein each image has a different label. The dataset exhibited an agreement rate of approximately 75% in about 80% of instances.

Table 3. List of datasets that encompass images typically sourced from search engines employing categorical values.

Dataset	Basic Emotion Labels
FER-2013 [67]	Angry, Disgust, Fear, Happy, Sadness, Surprise, Contempt
Google-FEC [102]	Amusement, Anger, Awe, Boredom, Concentration, Confusion, Contemplation, Contempt, Contentment, Desire, Disappointment, Disgust, Distress, Doubt, Ecstasy, Elation, Embarrassment, Fear, Interest, Love, Neutral, Pain, Pride, Realization, Relief, Sadness, Shame, Surprise, Sympathy, Triumph
EmotioNet [103]	Happy, Angry, Sad, Surprised, Fearful, Disgusted, Appalled, Awed, Angrily disgusted, Angrily surprised, Fearfully angry, Fearfully surprised, Happily disgusted, Happily surprised, Sadly angry, Sadly disgusted

EmotionNet [103] is a dataset distinguished by its automatic annotation of images using AUs with their corresponding intensities. This annotation process was achieved through the use of an algorithm. To enable automatic annotation, the algorithm was trained using three datasets that contained images manually labeled by human annotators. As a result, approximately one million images were annotated. After the images were assigned various AUs, they were categorized into 23 different emotional categories, as shown in Table 3.

4.3.3. Ordinal In-The-Wild Datasets

This type of dataset is characterized by images "in the wild" in which spontaneous expressions occur annotated in a continuous dimensional space.

AffectNet [41] is a dataset composed of images "in the wild" identified through the input of keywords in search engines and annotated according to a continuous dimensional space. A combination of different terms related to subjective states, ethnicity, gender, and age were used as keywords to identify images with emotional expressions. This resulted in 362 strings, which were then translated into five other languages. However, this last step did not produce the desired results when searching for images through search engines because of cross-cultural linguistic variations. Language is a highly articulated structure that has formed over time in relation to historical and cultural contexts [35]. Consequently, in order to identify emotion-related terms in AffectNet, subjects who were not native English speakers but possessed knowledge and fluency in the language were asked to translate the emotional labels. This resulted in the identification of a total of 1250 keywords, along with 450,000 images. To annotate the resulting images from the search, both a categorical model relating to eleven discrete categories and a two-dimensional continuous model referring to valence and arousal were employed, as shown in Table 4. The agreement was calculated only for a sample of the dataset, with only 36,000 images labeled by two annotators. The results show a relatively low inter-rater agreement of $P_o = 60.7\%$.

Table 4. List of datasets that encompass images typically sourced from search engines employing ordinal values.

Dataset	Basic Emotion Labels
AffectNet [41]	*Labels*: Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain, Non-face. *Dimensions*: Valence, Arousal
EMOTIC [104]	*Labels:* Affection, Anger, Annoyance, Anticipation, Aversion, Confidence, Disapproval, Disconnection, Disquietment, Doubt/Confusion, Embarrassment, Engagement, Esteem, Excitement, Fatigue, Fear, Happiness, Pain, Peace, Pleasure, Sadness, Sensitivity, Suffering, Surprise, Sympathy, Yearning. *Dimensions:* Valence, Arousal, Dominance
OMG-Emotion [105]	*Labels*: Surprise, Disgust, Happiness, Fear, Anger, Sadness. *Dimensions*: Valence, Arousal
EmoReact [106]	*Labels*: Happiness, Surprise, Disgust, Fear, Curiosity, Uncertainty, Excitement, Frustration, Exploration, Confusion, Anxiety, Attentiveness, Anger, Sadness, Embarrassment, Valence, Neutral. *Dimensions:* All emotions except valence are annotated on a 1–4 Likert scale
Aff-Wild [107]	*Dimensions:* Valence, Arousal
Aff-Wild2 [108]	*Labels*: Neutral, Anger, Disgust, Fear, Happiness, Sadness, Surprise. Partly annotated with 8 Action Units. *Dimensions:* Valence, Arousal

EMOTIC [104] is a database of images sourced from MSCOCO, Ade20k, and the Google search engine. A total of 18,316 images were annotated according to a categorical and three-dimensional model (VAD). Figure 6 illustrates samples from this dataset. The categories were identified from 400 vocabulary words relating to affective or psychological states. They were then grouped into clusters comprising terms linked by synonymic relationships. In this manner, 26 clusters were established, each pertaining to an emotion category, as illustrated in Table 4. The continuous annotation values in the dataset span a scale of 1 to 10. To assess the inter-rater reliability, the Fleiss Kappa coefficient was calculated, resulting in a mean value of $\kappa = 0.31$, with more than 50% of the images exhibiting a $\kappa > 0.31$.

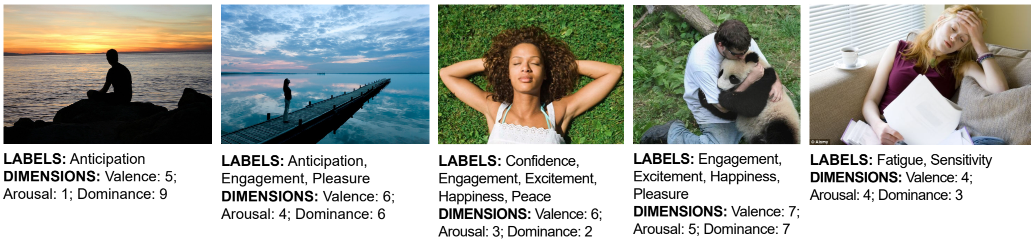

LABELS: Anticipation
DIMENSIONS: Valence: 5; Arousal: 1; Dominance: 9
LABELS: Anticipation, Engagement, Pleasure
DIMENSIONS: Valence: 6; Arousal: 4; Dominance: 6
LABELS: Confidence, Engagement, Excitement, Happiness, Peace
DIMENSIONS: Valence: 6; Arousal: 3; Dominance: 2
LABELS: Engagement, Excitement, Happiness, Pleasure
DIMENSIONS: Valence: 7; Arousal: 5; Dominance: 7
LABELS: Fatigue, Sensitivity
DIMENSIONS: Valence: 4; Arousal: 4; Dominance: 3

Figure 6. Samples of images and their related labels from the EMOTIC dataset [104].

The OMG-Emotion [105] dataset comprises 567 videos, with an average duration of approximately one minute. These videos were identified on various YouTube channels through the inclusion of keywords related to the term "monologue". The objective of the authors was to collect video content that presented a diverse range of emotions within the same context, with these emotions varying gradually over time. The videos were divided into clips (7371), each representing a distinct expression. These clips were then labeled by five independent annotators, who were asked to identify the six basic emotions recognized by Ekman (with the addition of "neutral") and the arousal/valence scale. This resulted in a two-dimensional model, as shown in Table 4. In addition, each annotator was provided with the context related to the clip of the complete video. Upon analysis of the distribution of labels, the authors noted that there is a greater concentration of labels around expressions that were annotated as neutral, as well as in relation to annotations related to neutral valence and calm arousal. However, despite this concentration, the values are distributed across the entire spectrum.

The EmoReact [106] dataset consists of video clips collected from a YouTube channel (React). The content of this channel consists mainly of children (12–14 years old) reacting to different events and showing different emotions. Thirty-seven subjects were selected; then, the videos were divided into clips with an average length of 5 s, picking only those with a minimal length of 3 s. In this way, 1254 clips were obtained. These were then annotated as shown in Table 4. The dimensional element in this dataset is expressed by the presence of annotations on a 1–4 Likert scale for each emotional label, where 1 indicates the absence of the emotion and 4 corresponds the intense presence of the emotion; this was applied to all categories except valence, which was annotated on a 1–7 scale, where 1 indicates the maximum negativity of emotion and 7 represents the maximum positivity of emotion. The inter-judge agreement (Krippendorff's α) reported by the authors was calculated, showing an agreement of $\alpha \in [-0.16, 0.64]$.

The Aff-Wild [107] dataset contains 298 videos, mainly sourced from the YouTube search engine. The clips were annotated with respect to valence and arousal. For the annotation task, the authors developed their application, allowing users to evaluate arousal and valence separately through the use of a joystick. Following the validation of the annotations, every judge was asked to re-watch the video to check the accordance between it and the labels. Cross-correlations were calculated between all annotators for each video. Furthermore, the correlation between the annotations and the tracked facial landmarks was determined. This resulted in the establishment of a ranking of the annotators' correlations for each video. Subsequently, two additional experts reviewed all videos and selected the annotations that were mostly correlated. The mean of the selected annotations, which were highly correlated and received positive evaluations from the experts, was then computed.

Aff-Wild2 [108] is a collection of 545 videos sourced from YouTube. It represents an extension of the previously mentioned Aff-Wild [107]. The videos were validated using basic emotion labels, valence/arousal, and AUs, as reported in Table 4.

5. Face Emotion Recognition Concerns

In this section, we present the various concerns that have emerged within the literature on the topic of FER, categorizing them into three distinct perspectives. In recent years, there has been a vast amount of discussion around different aspects or challenges of automatic emotion recognition [16]. Indeed, a significant proportion of the literature challenges the very concept of automatic emotion recognition, arguing that the psychological models on which it is based are not valid or scientifically sound [2,6,109]. Secondly, a number of scholars contend that, regardless of the feasibility of emotion recognition, the use of this technology could be dangerous and potentially harmful to human rights [69,110–112]. In a third instance, a portion of authors maintain that, even if it is assumed that the psychological models on which FER is based are perfectly valid and these systems are not harmful to individuals, there persists a substantial problem, namely the lack of reliability of FER ground truth [16]. Several studies have attempted to systematize different aspects of FER, including its potential ethical harms and the psychological foundations underlying the technology. For instance, Amelia Katirai [69] conducted a systematic review of 43 articles concerning ethical considerations with respect to FER. Moreover, Soumya Ranjan Mohanta and Karan Veer [81] identified different challenges associated with FER, which consist of capturing the context of emotions, lighting issues, racial differences, and variations in facial expressions. On the other hand, various studies have explored the claims of BET and the relationship between facial expressions and emotions. For example, Juan I. Durán and José-Miguel Fernández-Dols [7] conducted a meta-analysis to test whether whole facial expressions co-occur with basic emotions and whether parts of facial expressions align with specific emotions. They also examined the hypothesis that a linear relationship exists between the intensity of emotional experience and the intensity of facial expression.

Their findings indicated a low co-occurrence between facial expressions and the subjective experience of emotion, both for whole and partial facial expressions. Moreover, they observed significant variability in the correlations of emotional and facial expression intensities. Similarly, Lisa Feldman-Barrett et al. [5] surveyed examples of psychological scientific evidence to test the assumptions underlying the "common view" of emotions. They specifically examined the typical experimental design and conclusions about the existence of distinct facial expressions or the universality of recognition. However, the reliability of FER's ground-truth data and its impact on accuracy are often overlooked [16]. Additionally, to the best of our knowledge, the various perspectives on FER have rarely been organized in a single comprehensive framework.

In the following section, we propose a three-layer conceptual categorization to better address the diverse issues surrounding FER technology and clarify the ongoing debates within the scientific community. Notably, we examine concerns related to the psychological foundations of FER, its potential ethical harms, and the reliability of its ground truth, underlining how different levels of analysis can impact various aspects of FER technology.

5.1. Challenging Emotion Fingerprints

Despite the existence of a broader spectrum of models, FER is mainly based on the oversimplistic emotional theories of Ekman and others [2,28]. Thus, fundamentally, it rests upon the aforementioned belief that there exists a direct correspondence between basic emotions (i.e., private feelings) and facial expressions [5,28,111]. In other words, the underlying assumptions of emotion recognition systems are derived from only one part of the psychological tradition relating to the study of emotions [28].

This culturally entrenched belief stemming from the "universality hypothesis" posits the existence of emotional "fingerprints", which supports the notion that internal and private states can be measured. Universal recognition of emotions can only be achieved if they are produced universally; hence, facial expressions must be reliable, univocal, and diagnostically distinctive marks to enable their categorization. Within this perspective, facial expressions should be considered robust enough to account for basic emotion classification [2].

This conviction is not only carried forward by scientific psychological research, but it is also reinforced by various cultural products that iterate the "common view" of emotions [5]. For instance, media content, such as movies or television series, usually represents precise face patterns as straightforwardly portraying specific emotions, which are consequently universally recognized [5]. That is the case, for example, of the movie "Inside Out", in which five basic emotions (anger, disgust, fear, sadness, and joy) are represented by fictional characters and for which Ekman offered scientific support [113].

Nevertheless, the study of emotions is much more complex than this "common view" [5,6], and in the literature, there is not a complete agreement on the robustness of emotion labels in unambiguously determining subjective states [1,6,17,18] to the extent that some argue that the current framework of study is creating a barrier for their in-depth exploration [2].

Barrett, for instance, due to the inconsistencies that happen in emotion recognition tasks, criticized both the application in fields such as AI and its theoretical foundations [5], also arguing that

> "*When it comes to emotion, a face doesn't speak for itself. In fact, the poses of the basic emotion method were not discovered by observing faces in the real world. Scientists stipulated those facial poses, inspired by Darwin's book, and asked actors to portray them. And now these faces are simply assumed to be the universal expressions of emotion* [2]."

This perspective and strong opposition to the notion of emotions as discrete and measurable entities [6] stem from a series of experiments that the author conducted while investigating the origins of low self esteem and its connection to anxiety or depression. Failing to replicate known phenomena, upon closer examination, a consistent anomaly was discovered; participants experienced difficulty in distinguishing between feelings of anxiety and depression, indicating a struggle in discerning between these emotions. In subsequent experiments, subjects were asked to monitor and annotate their emotional experiences. The results showed that participants employed shared terminology, such as “angry”, “sad”, and “afraid”, to describe their feelings; however, these words did not always indicate the same emotional states. This pattern also occurred with positive emotions, such as “happiness”, “calmness”, and “pride”. After conducting a thorough examination with more than seven hundred American participants, to the author found a significant variation in how individuals describe and feel their emotional experiences [2].

Contrary to Ekman’s assertion, the idea that a facial expression unambiguously expresses a precise emotion is not evident [1,2,5,6,30]. The proponents of this standpoint affirm that facial expressions alone do not offer a clear indication of emotions, as they are dependent on the perceiver. One set of facial expressions can be linked to different emotion categories, and multiple sets can be connected to a single emotion, standing in a many-to-many correspondence. More simply, a face alone does not convey emotions clearly [1,2,6,109].

Barrett et al. [5], identified the following three key limitations of the scientific literature that are associated with the misinterpretation of emotions: limited reliability, lack of specificity, and limited generalizability. The initial concept pertains to how emotions belonging to a single category do not exhibit homogeneous facial expressions nor they are universally recognized through a standardized set of facial movements. The second point highlights that there is no univocal mapping between a specific arrangement of facial movements and occurrences of a particular emotional category. The third refers to the influences of context and culture on emotional expression.

According to the latter point, contextual and cultural information assumes central relevance; in particular, body posture, gestures, social situation, and culture are considered sources of information as relevant as facial expressions [5]. Specifically, the cultural component has been under the lens of several scientific analyses [10]. The expression and perception of emotions in humans, as well as the discernment of pertinent information to distinguish one emotion from another, are profoundly influenced by cultural context [5,10]. For instance, Matsumoto [114] thoroughly documented how cultural differences influence the rate of correct recognition of a specific emotion. Moreover, he argued that recognition accuracy varies between cultures. Individuals often interpret the emotions of their ethnic group more accurately than those of different cultural backgrounds. The researcher’s experiments involved American and Japanese college students asked to identify six of Ekman’s universally recognized emotions from a set of 48 photographs. These photos depicted expressions from two male and two female individuals of both Japanese and Caucasian descent. The experimental results demonstrated that Americans exhibited a heightened precision in discerning emotions such as anger, disgust, fear, and sadness. Furthermore, this accuracy was consistent across both genders and unaffected by cultural factors. In contrast, Japanese participants showed a greater facility in recognizing these emotions when expressed by females compared to males.

Cultural values play a key role in determining coordination and organization within a society, providing a system of information shared by its members. Indeed, the values concerning interpersonal relationships (e.g., individualistic/collectivistic) and the values related to emotions are crucial in determining emotional regulation norms [115]. In particular, the values pertaining to emotions provide rules or guidelines for their desirability within specific social contexts [115,116]. Consequently, the exhibition of emotions is governed by “display rules”, namely the learned and culturally given differences that determine how each emotion is expressed in various social situations [9,116]. For instance, emotion suppression is often associated with negative social consequences when exhibiting a certain emotion in social contexts [117]. Cross-cultural experimental research conducted by Emily A. Butler et al. [117] has supported this view, revealing that women who primarily adhere to European values tend to engage in less habitual emotional suppression than their bicultural European–Asian counterparts. Moreover, among women with strong European values, this suppression is often associated with self-protective objectives and an increase in negative emotions. In contrast, bicultural women exhibit the opposite trend, where suppression correlates with less self-protective goals and reduced negative emotions.

As a result of the points underlined, the theory of "constructed emotions" [2,118] proposes a change of paradigm to leave essentialism in emotions and to accept variability as the norm rather than the exception. This view provides a multi-level, constructionist view of the brain's basis for emotion, aligning with computational and evolutionary biology. The theory challenges traditional views by asserting that emotion categories lack distinct neural bases and emphasizes the importance of neural ensembles over individual neurons. Accordingly, the possibility of measuring an emotion exclusively by means of indirect data, such as facial movements or physiological alterations, is deemed to be untenable.

This perspective on emotions and the concerns identified within BET take FER apart on the basis of its grounds, namely that it undermines the validity of the psychological models on which it is based. Hence, if emotions cannot be measured and assigned to a corresponding label, then this impossibility is reflected in FER systems. This technology, consisting essentially of ML algorithms trained on large datasets, relies on images annotated by human voters who have no prior knowledge of the subjects in the images [16], in compliance with a precise psychological model [28]. According to this view, the design of emotion recognition tools is influenced by oversimplistic models, typically Ekman's six basic emotions [28]. In simpler words,

> *With machine learning, in particular, we often see metrics being used to make decisions—not because they're reliable, but simply because they can be measured* [8]."

Besides the fact that ML algorithms are trained on reductionist models of emotions, they also neglect to consider variations in perception that are given by context, culture, and the distinction between fake and genuine emotions [111]. Indeed, many FER datasets are based on pictures portraying actors simulating a specific expression (according to basic emotion labels) in laboratory settings and in the absence of the context in which they originated [5,16,93]. Accordingly,

> "*Emotions are complex, and they develop and change in relation to our families, friends, cultures, and histories, all the manifold contexts that live outside of the AI frame* [111]."

Moreover, in real-life scenarios, normative implications can be derived from the outputs of these systems. Therefore, the ethical responsibilities of human beings also depend on the theory of emotion being considered [28], encouraging discussions for precise regulation.

5.2. Ethical Concerns Associated with Face Emotion Recognition

In recent years, there has been an increase in awareness regarding the ethical aspects surrounding the AI scenario [29,69]. As previously mentioned in Section 4, some FER applications have raised questions among various scholars about the possible risks and unfair employment of this technology. These discussions examine whether these systems have the potential to cause harm to individuals, particularly those who belong to minority groups, or those who are considered at-risk, such as children [50,111,119]. As emotion recognition applications experience rapid expansion, becoming a lucrative market [69,112], the concerns with respect to the possible harms of potential biases and the lack of transparent models also are increasing in association with this specific technology [69,111]. For instance, Katirai [69] identified three key risk areas, namely the danger of biased and unfair outcomes, the sensitivity of emotion to data or privacy harms, and the use of emotion recognition in particular and consequential contexts.

Stemming from the underpinning of FER by the emotion model [28,69] and the subjectivity of the labeling process [77], biases and unfairness have the potential to engender discriminatory practices in the deployment of FER systems. Despite being used often in an interchangeable way, fairness reflects the subjective judgment of how a construct is measured and justified when applied in decision making, while biases represent systematic errors that disproportionately impact evaluations across different groups [120]. Indeed, psychological and cross-cultural literature has highlighted not only differences between cultures in emotion regulation but also differences in emotion classification based on race [121], which can be perpetuated through the labeling process of the FER ground truth [16,28]. For instance, experiments conducted by Paul B. Hutchings and Geoffrey Haddock [122] demonstrated that participants with high implicit prejudice were more likely to categorize racially ambiguous angry faces as Black and to rate the intensity of the anger as greater when the faces were categorized as Black compared to White. Eugenia Kim et al. argued that although racial biases have been widely studied in facial FER, age-related biases remain largely overlooked [123]. Their research highlights how facial morphology changes with age, affecting emotion recognition accuracy, and identifies a gap in current FER benchmarking protocols, which often fail to consider age as a significant factor. Moreover, various authors have argued that the datasets on which FER algorithms are trained are unevenly distributed in terms of ethnicities, gender, and age, resulting in numerous biases, consequently impacting these minorities [124–126]. This calls for more inclusive, intersectional algorithm development and evaluation practices.

Additionally, many scholars are concerned with the sensitivity of emotions if collected as data [28]. In particular, the limits of expressions in being defined as private or public are variable [127], resulting in the collection of this type of information being considered an invasion of the individual sphere. Intelligent algorithms are continuously fed with biometric inputs in an attempt to enhance people's experiences. However, this constant prediction of our behaviors puts indispensable rights at risk [128], becoming a new source of power [111]. The concern for such values assumes a pivotal relevance as they are employed in more and more application contexts, namely airport surveillance [129], car driving safety [53], and in support of autistic individuals [130]. For instance, applications in which FER technology is declared to improve the quality of life of people with disabilities are particularly critical but result in the exploitation of these conditions to push rhetoric aimed at promoting surveillance capitalism. This is the case, for instance, of Affectiva's Affdex, in which autism is used as the rationale for rendering emotions computable and for advancing commercial emotion AI [131].

As a consequence, various authors have proposed guidelines, frameworks, and mitigation strategies to better evaluate the risks and harms associated with FER technology. For example, Javier Hernandez et al. [29] developed recommendations for assessing and minimizing risks related to emotion recognition, which include responsible communication, informed consent, contextual calibration, and comprehensive contingency planning. Additionally, Andrew McStay and Pamela Pavliscak [110] created an ethical checklist for the use of emotional AI, emphasizing, for instance, the importance of recognizing the lack of global agreement on emotions and ensuring that these technologies provide a clear benefit to the user. In addition to these guidelines, a few regulatory frameworks have been proposed to address the legal and ethical concerns surrounding FER technology. These frameworks aim to ensure compliance with privacy laws and data protection standards while promoting responsible deployment. One notable example is the European Union (EU) AI Act, the world's first comprehensive AI legislation, which explicitly addresses FER [132], as examined in greater detail in the subsequent paragraph. However, regulations concerning AI are rapidly evolving. For instance, the National Institute of Standards and Technology (NIST)'s AI Risk Management Framework provides guidance for organizations, governments, the private sector, and civil society on AI use. However, although it acknowledges the risks associated with biometric data, it does not specifically regulate emotion recognition [133].

What Is the Position of the EU?

The debate surrounding the regulation of emotion recognition technology has seen diverse perspectives, especially in the European context. Several civil society organizations have advocated for a comprehensive prohibition on the use of this technology. This view was initially adopted by the European Parliament in the early drafts of the EU AI

Act, which instituted a ban on emotion recognition across the following four key specific domains: educational settings, workplaces, law enforcement, and immigration processes. Consequently, the Act introduced exceptions allowing for emotion recognition only in particular contexts like health care and safety. While these exceptions aim to enable beneficial applications, their imprecise boundaries also risk unintended consequences. Moreover, subsequent revisions of the Act also led to the exclusion of law enforcement and immigration control from the ban, which has been confirmed in the final draft. This adjustment warrants thoughtful examination, given its differential impact on vulnerable communities.

Specifically, the perspective embraced in the June 2023 amendments seems to recognize the subjective nature of emotions. Indeed, Article 26c affirms that

> "*The key shortcomings of such technologies, are the limited reliability (emotion categories are neither reliably expressed through nor unequivocally associated with, a common set of physical or physiological movements), the lack of specificity (physical or physiological expressions do not perfectly match emotion categories) and the limited generalisability (the effects of context and culture are not sufficiently considered)*" [134].

Moreover, the final draft reiterates these statements and also supports cross-cultural differences in emotion recognition [49]. Nevertheless, despite the concerns about the scientific basis of emotion recognition, the 2024 final draft of the AI Act, providing exceptions rather than an overall ban, leaves room for employment. Deployers are required to properly inform individuals about the system's operation and adhere to pertinent data regulations. However, this provision excludes AI systems used for biometric categorization and emotion recognition for the detection, prevention, and investigation of criminal offenses. In addition, emotion recognition systems are only prohibited in work and school environments, with exceptions for safety and medical reasons [49].

The urgency to protect essential rights becomes increasingly relevant if we take into account that fact that emotions are instances of biometric data that can be used to profile individuals [128], resulting in the previously mentioned surveillance capitalism [23]. Indeed, it

> "*Claims human experience as free raw material for translation into behavioral data*" [23].

In consideration of the above, the EU's position appears to be somewhat contradictory. While it criticizes the scientific soundness of emotion recognition, it also allows exceptions for the use of biometric data in particular circumstances. Indeed, the employment of remote surveillance systems creates an inevitable imbalance of power between those who control and those who are controlled; people's voices, faces, emotions, and bodies are used as raw data that feed an algorithm that invades private life [23,111].

5.3. On the Dubious Reliability of Face Emotion Recognition

The third area of concern involves the reliability of FER. While prior discussions have highlighted the lack of consensus on the measurement of emotions from indirect signals and the potential for harmful outcomes in their application, in this section, we assume, for the sake of argument, that FER systems are based on valid and universally accepted psychological models and that their real-world applications are entirely ethical, transparent, and fair. Even under these ideal conditions, a critical issue remains, namely whether these systems can reliably recognize emotions. Automatic emotion recognition consists of a categorization process in which, through probabilistic measurement, the dominant emotion is detected. From a technical point of view, this technology can be related to two important concepts, namely validity and reliability [16]. In particular, FER validity, that is, the accuracy of emotion recognition, is related to the underlying theoretical psychological foundations [5]. In other words, recognition is considered valid if it identifies the intended emotional state [16]. In contrast, reliability refers to consistency across different times, places, subjects, and experimental conditions [135,136]. More specifically, in the FER literature, it pertains to how persistent different subjects are in assigning a particular emotional label to a person's face depicting an expression [16], namely inter-rater agreement [135]. Because FER datasets have a vast number of images, the labeling task is typically undertaken by multiple raters, who are given different expression images "in the wild" (i.e., pictures downloaded from various Internet sources) [93]. Consequently, these datasets are susceptible to "annotation bias", which refers to systematic errors in the labeling process affecting how emotions are labeled [137]. Additionally, since FER depends on ML algorithms, its reliability is inherently tied to its ground truth and impacts model accuracy [16,51].

For a more comprehensive grasp of the concept of reliability and its relevance to emotion annotation, it can be helpful to draw an analogy with the task of lie detection. The attribution of a personal and internal state, such as intent to lie or an emotion, is based on an inference from external data to inaccessible information, such as heart rate/blood pressure, respiration, and skin conductance in the case of a polygraph [138] or facial expressions in the instance of FER [28]. As a result, we are faced with an intrinsically subjective task, which is based on interpretative judgments [6,139]. Lie detection tools, often portrayed as "magic mind-readers" [135], have been long debated, putting int question their validity and reliability [135,138,140]. Critics argue that physiological changes registered by polygraph may be derived from aspects beyond deception, since they are determined by plausible anxiety or fear, which can be mislabeled [135,141]. These discussions give rise to controversies with respect to its application, as it is often considered non-sufficiently reliable to be accounted as evidence [140] (e.g., United States v. Sheffer in 1998 [142]). On the other hand, although emotion judgments are regarded as subjective and perceiver-dependent in various psychological contributions [5], this is rarely linked to the potential influence on the creation of FER technologies [16].

Hence, the question of whether emotion recognition systems can be considered sufficiently reliable arises. However, before attempting to provide a response, it is necessary to discuss a further point, namely what levels of reliability are considered "high enough". Various metrics are used to assess inter-rater reliability. These methods include the simple percentage agreement, as well as more robust measures, such as Krippendorff's α [143]. Moreover, Krippendorff suggests that a reliability measure of $\alpha \geq 0.800$ is customary for agreement (adequacy threshold), while values of $\alpha \geq 0.667$ are acceptable. Values of $\alpha \leq 0.667$ (unacceptability threshold) are considered too low to detect agreement and should be discarded [16,143]. However, in general, the degree of subjectivity and disagreement increases as the value of the reliability measure decreases [16].

Returning to the question concerning sufficient reliability, from the review of the datasets presented in Section 4.3, what emerged is that between them, all the reliability scores are below 0.800, except for the Radboud dataset [101]. Moreover, not all the datasets report their reliability score. Focusing on the settings with which which the images were captured, the initially developed datasets typically comprise pictures collected in a controlled environment—commonly a laboratory setting [41]. In these situations, facial expressions are elicited through stimuli or subjects are asked to exhibit facial behavior voluntarily. As previously discussed, there are numerous differences between genuine and simulated expressions [35]. Indeed, the latter are often exaggerated, accentuating the differences between expressions, and do not exhibit occlusions [144]. This results in greater ease of classification and higher agreement values [144]. Considering the model employed for the labeling of emotion, the prevalence of the analyzed datasets presents expressions that were annotated according to a categorical model, which typically includes the six emotions identified by Ekman as basic [14]. However, there are databases with a considerably higher number of labels. Furthermore, several datasets were annotated by FACS [48] experts through the use of AUs (e.g., Radboud Faces Database [101]). A small number of datasets were annotated with ordinal values, despite the fact that they allow for the measurement of the intensity of a given emotion, thereby capturing variations in more detail [45].

Federico Cabitza et al. [16] conducted a study to evaluate the reliability of FER ground-truth data by organizing a user study with two annotation experiments designed to test the following three research questions:

- Is the inter-rater reliability of the FER ground truth sufficient to support reliable research and analysis?
- Does providing some sort of contextual information have any effect on the reliability of the ground truth?
- Is the intra-rater reliability of the FER ground truth high enough?

In the first experiment, inter-rater reliability was assessed through an online questionnaire in which participants annotated 30 genuine (non-posed) pictures of facial expressions using Ekman's basic emotion labels. Additionally, for each picture, participants rated the perceived intensity of each emotion on a scale from 1 to 5, indicating the degree to which they felt the emotion was present. Figure 7 and Figure 8 show respectively samples from their dataset and a screenshot of their questionnaire. The tested subjects were randomly divided into the following two groups: a no-context group (shown only randomly ordered pictures) and a context group (shown the pictures and the videos from which they were extracted) in order to assess the second hypothesis. In the second experiment, they evaluated intra-rater reliability by testing different participants with a slightly modified version of the first experiment (context group) in which 5 of the 30 pictures were repeated. In both the first and second experiments, intra-rater and inter-rater reliability was evaluated through Krippendorff's α [143] considering the threshold mentioned above. The following three types of representations were considered for each rating: label-based (categorical model), distribution-based (each rater's judgment is represented as a percentage), and ordinal-based (each emotion judgment is represented as a list of reported emotional intensities) representations.

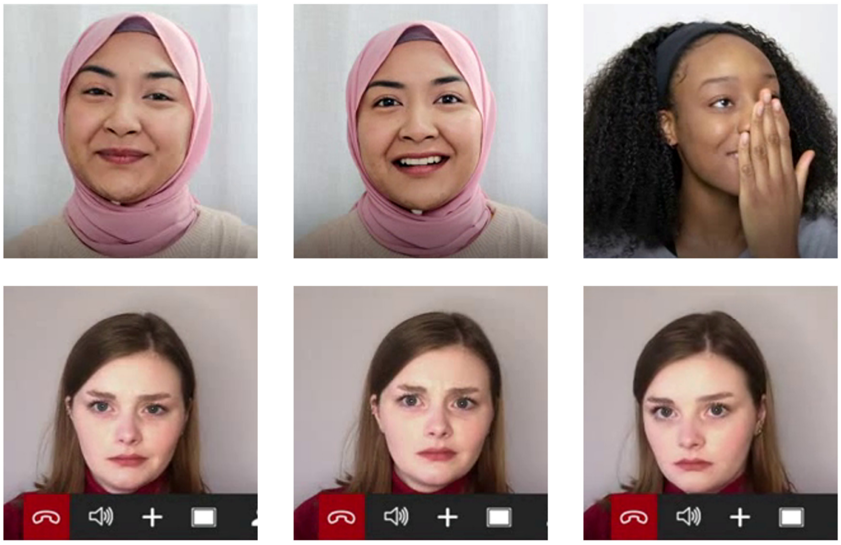
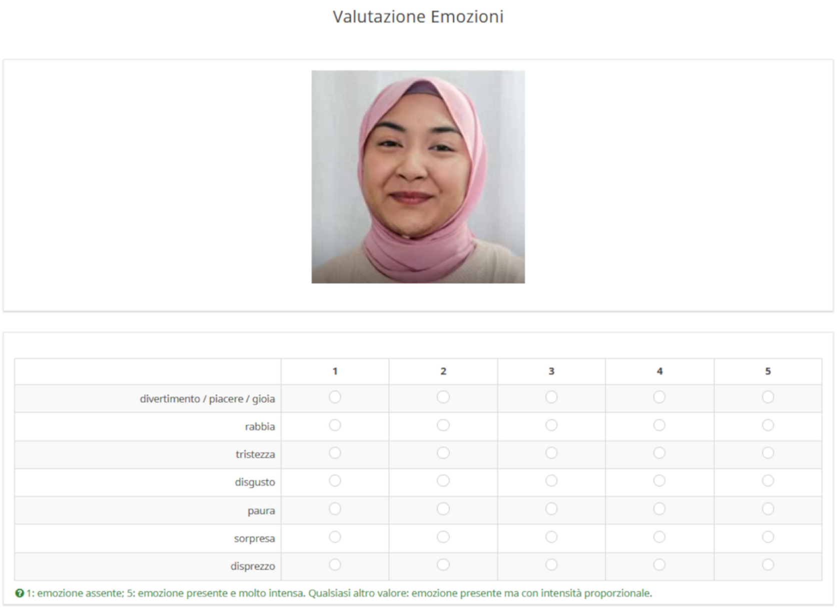


Figure 7. Six images from the Cabitza et al. study depicting three subjects whose emotions were assessed by study raters. The top images had the highest agreement (easiest to interpret), while the bottom images had the lowest agreement (hardest to detect) [16].

In the first experiment, all inter-rater reliability values were significantly below the adequacy threshold. Additionally, all values, except for the ordinal representations of "anger" (in both groups) and "fear" (in the context group), fell significantly below the unacceptability threshold. Furthermore, the reliability values for the context group surpassed those of the no-context group. In the second experiment, the reliability values for the multi-label and distribution-based representations were significantly below the unacceptability threshold. Specifically, the reliability values for the emotions of "sadness", "anger", and "fear" exceeded the unacceptability threshold. In contrast, all other reliability scores were below the adequacy threshold. Given these considerations, the authors concluded that systems based on facial expressions lack sufficient levels of ground-truth reliability, as humans do not agree enough in the emotion annotation task, highlighting the subjective nature of emotion recognition and the resulting risk of bias. Furthermore, these findings are consistent with the review of existing datasets, clearly indicating that the FER community is grappling with a notable issue of low ground-truth reliability. Therefore, in addition to ethical and foundational concerns, these considerations should serve as a compelling limitation of the deployment of automated emotion recognition systems [16].



Valutazione Emozioni

	1	2	3	4	5
divertimento / piacere / gioia	○	○	○	○	○
rabbia	○	○	○	○	○
tristezza	○	○	○	○	○
disgusto	○	○	○	○	○
paura	○	○	○	○	○
sorpresa	○	○	○	○	○
disprezzo	○	○	○	○	○

? 1: emozione assente; 5: emozione presente e molto intensa. Qualsiasi altro valore: emozione presente ma con intensità proporzionale.

Figure 8. Instance of an annotation page extracted from Cabitza et al.'s questionnaire [16].

6. Discussion

The debate surrounding FER technology is both intricate and multifaceted, reflecting the complexity of human emotions and the challenges inherent in their accurate recognition and interpretation. There is currently no universally accepted theory of emotions, nor is there a consensus on a precise definition of emotions within the scientific community. While various models of emotion exist, FER systems are predominantly based on the motivational theories of Ekman and others, which posit a direct link between basic emotions and facial expressions. This assumption, which is deeply rooted in Western culture, implies that emotions have distinct and measurable facial "fingerprints". However, this perspective has been increasingly challenged by critics who argue that emotions are far more nuanced, context-dependent, and difficult to categorize into discrete entities. These limitations are also mirrored in the low reliability and annotation biases observed in FER ground-truth data. Moreover, the uncertainty of psychological foundation of FER becomes increasingly significant when the technology is applied in contexts that pose potential threats to individual privacy and civil rights.

Our review has revealed three critical areas of concern surrounding FER technology, namely its foundational theories, its ethical implications, and its technical reliability.

- First, the reliance of FER systems on oversimplified models of emotions raises concerns about the validity of their psychological foundations. These systems often overlook the influences of context, culture, and the distinction between genuine and simulated emotions.
- Secondly, the potential for racial biases and the perpetuation of these stereotypes through AI systems pose significant ethical challenges. Studies have shown that FER systems can exhibit variations in emotional interpretation influenced by an individual's race, leading to troubling outcomes, especially in surveillance and law

enforcement settings. As a consequence, there is growing awareness regarding the ethical aspects surrounding the FER scenario, questioning whether these technologies have the potential to cause harm to individuals, particularly those who belong to minority groups or those who are considered at-risk, such as children.

- Finally, even if assuming that the first two requisites are entirely met and non-problematic, meaning that emotion recognition systems are based on valid psychological models and that their application poses no ethical risks, there is a third point to be touched upon, namely the reliability of the ground truth utilized by these systems. In other words, the ground truth on which FER is built is unreliable due to insufficient agreement values and biased datasets.

Based on these considerations, it can be stated that the current state of FER technology is still a long way from being a reliable and ethical tool for emotion recognition. Subsequent studies should aim at the construction of more culturally sensitive and contextually grounded models of emotion that do not rely on the paradigm of universality. This includes expanding the number of physiological and contextual cues used to enhance the identification of emotions. Nevertheless, this article does not seek to prescribe specific frameworks, policies, or political directives. Instead, it aims to serve as an inclusive and informative resource for practitioners, policymakers, and legislators involved in addressing the various challenges related to FER. However, a key recommendation that can be derived from our analysis is the necessity of explicitly reporting the accuracy rates of FER technologies, along with a thorough evaluation of the datasets used in such technologies, including in terms of their representativeness and reliability. This recommendation aligns with the transparency obligations mandated by regulatory frameworks such as the AI Act [49] and the General Data Protection Regulation (GDPR) [145], which emphasize the importance of disclosing the sources of data and performance metrics to ensure accountability and trustworthiness in AI systems.

Limitations

The field of emotion recognition and its associated regulations is rapidly evolving. Both technical advancements and regulatory developments are ongoing, which may impact the relevance and applicability of our findings over time. As such, this review should be considered a reflection of the current state of the field, with the understanding that future changes could alter the dynamics we have discussed. Furthermore, despite our review covering a broad spectrum of topics, the scope of our analysis was restricted to the evaluation of FER technologies, focusing primarily on facial expressions as a proxy for emotional states. While in this paper, FER was widely studied, alternative methods for emotion recognition, such as those involving biosignals or other physiological measures, were not examined in depth. Literature focusing on FER algorithms and applications was not a primary focus of this article, so we refer the reader to more in-depth surveys [78–81,89]. Moreover, although our article presents various frameworks, guidelines, and checklists, its primary purpose is not to prescribe specific frameworks but, rather, to provide a thorough analysis. However, future research can better address the various nuances of the FER discourse using the comprehensive classification we have provided.

7. Conclusions

In this paper, we provided a comprehensive overview of the current debate on FER technology, framing it around three pivotal perspectives that shape the discourse. By synthesizing these viewpoints, we aim to contribute to a more informed and balanced debate about the future of FER systems. We argue that addressing these concerns requires interdisciplinary research and dialogue, supported by the development of comprehensive frameworks that move beyond isolated analyses of individual aspects of FER. Indeed, we believe it is necessary to consider three different levels of concern or critical areas to properly address its foundational, ethical, and technical implications.

We argue that the formulation of guidelines and regulations necessitates not only the identification of high-risk applications and uses but also an evaluation of the theoretical foundations, training datasets, and reliability of the ground truth upon which these systems are based. While current European regulatory frameworks acknowledge the limitations of emotion recognition through the use of proxy data, they often fall short of addressing the deeper issues highlighted in the psychological literature, such as the challenges of measuring emotions accurately and reliably. Importantly, these frameworks allow for certain applications, particularly in law enforcement, without fully accounting for the broader ethical and technical implications. Therefore, we advocate for a multi-layered analytical approach to regulation that considers the interconnectedness of foundational, ethical, and technical concerns. This stratified approach, as illustrated in Table 5, would enable researchers and policymakers to better identify potential pitfalls of each critical area of FER research and applications, thereby ensuring a more reliable and equitable deployment of these systems. Notably, negative outcomes stem not only from inappropriate or harmful uses but also from the fragile foundations and low reliability of ground-truth data. These factors significantly impact the uncertainty of emotion recognition systems and must be considered by researchers.

Table 5. Summary of the three main perspectives on the FER debate.

Area of Criticism	
Psychological Foundations	The psychological foundations on which FER technology is based are not uniformly accepted and suffer from theoretical ambiguities. Emotions are not considered measurable "entities" and do not stand in a univocal relation with expressions.
Ethical Implications	Emotions can be considered soft biometric data, feeding surveillance capitalism. Their employment in sensitive scenarios can be potentially harmful to essential human rights.
Reliability Issues	FER ground truth is considered unreliable and datasets may replicate annotation biases. Human beings do not agree sufficiently in the emotion annotation task.

In summary, our review

- Surveyed the complexity of human emotions, from the early philosophical theories to the major psychological models;
- Provided an overview of how FER works and its common applications, describing the most frequently used datasets;
- Identified three different areas of concern, developing a taxonomy for the analysis of FER issues, namely the psychological foundations, possible negative ethical outcomes, and reliability of these systems;
- Advocated for a multi-layered approach that focuses on how the various areas of criticism are interconnected in order to help researchers and policymakers better address the implications of FER.
- Emphasized the need for interdisciplinary research and careful regulation to improve the reliability, ethical responsibility, and effectiveness of FER systems, particularly in safeguarding marginalized and vulnerable populations.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank Antonio Emanuele Cinà and Marcello Pelillo for their precocious support. This work has been done while Martina Mattioli was enrolled as a student in the Italian National PhD Program in Artificial Intelligence for Industry 4.0.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Gendron, M.; Roberson, D.; van der Vyver, J.M.; Barrett, L.F. Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion* **2014**, *14*, 251. [CrossRef] [PubMed]
2. Barrett, L.F. *How Emotions Are Made: The Secret Life of the Brain*; Houghton Mifflin Harcourt: Boston, MA, USA, 2017.
3. Gates, K.A. *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*; NYU Press: New York, NY, USA, 2011.
4. Berry, J.W.; Poortinga, Y.H.; Pandey, J. *Handbook of Cross-Cultural Psychology: Basic Processes and Human Development*; John Berry: Boston, MA, USA, 1997; Volume 2.
5. Barrett, L.F.; Adolphs, R.; Marsella, S.; Martinez, A.M.; Pollak, S.D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol. Sci. Public Interest* **2019**, *20*, 1–68. [CrossRef] [PubMed]
6. Barrett, L.F. Solving the emotion paradox: Categorization and the experience of emotion. *Personal. Soc. Psychol. Rev.* **2006**, *10*, 20–46. [CrossRef] [PubMed]
7. Durán, J.I.; Fernández-Dols, J.M. Do emotions result in their predicted facial expressions? A meta-analysis of studies on the co-occurrence of expression and emotion. *Emotion* **2021**, *21*, 1550. [CrossRef] [PubMed]
8. Vincent, J. Emotion Recognition Can't be Trusted. 2019. Available online: https://www.theverge.com/2019/7/25/8929793/emotion-recognition-analysis-ai-machine-learning-facial-expression-review (accessed on 7 September 2024).
9. Matsumoto, D. Ethnic differences in affect intensity, emotion judgments, display rule attitudes, and self-reported emotional expression in an American sample. *Motiv. Emot.* **1993**, *17*, 107–123. [CrossRef]
10. Barrett, L.F.; Mesquita, B.; Gendron, M. Context in emotion perception. *Curr. Dir. Psychol. Sci.* **2011**, *20*, 286–290. [CrossRef]
11. Hofstede, G. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*; Sage: Melbourne, VIC, Australia, 2001.
12. Matsumoto, D. Cultural influences on the perception of emotion. *J. Cross-Cult. Psychol.* **1989**, *20*, 92–105. [CrossRef]
13. Russell, J.A. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychol. Bull.* **1994**, *115*, 102. [CrossRef]
14. Ekman, P.; Friesen, W.V. Constants across cultures in the face and emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124–129. [CrossRef]
15. Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 2000.
16. Cabitza, F.; Campagner, A.; Mattioli, M. The unbearable (technical) unreliability of automated facial emotion recognition. *Big Data Soc.* **2022**, *9*, 20539517221129549. [CrossRef]
17. Russell, J.A. Emotion, core affect, and psychological construction. *Cogn. Emot.* **2009**, *23*, 1259–1283. [CrossRef]
18. LeDoux, J.E.; Hofmann, S.G. The subjective experience of emotion: A fearful view. *Curr. Opin. Behav. Sci.* **2018**, *19*, 67–72. [CrossRef]
19. Hugenberg, K.; Bodenhausen, G.V. Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychol. Sci.* **2004**, *15*, 342–345. [CrossRef] [PubMed]
20. Hugenberg, K.; Bodenhausen, G.V. Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychol. Sci.* **2003**, *14*, 640–643. [CrossRef] [PubMed]
21. Halberstadt, A.G.; Castro, V.L.; Chu, Q.; Lozada, F.T.; Sims, C.M. Preservice teachers' racialized emotion recognition, anger bias, and hostility attributions. *Contemp. Educ. Psychol.* **2018**, *54*, 125–138. [CrossRef]
22. Halberstadt, A.G.; Cooke, A.N.; Garner, P.W.; Hughes, S.A.; Oertwig, D.; Neupert, S.D. Racialized emotion recognition accuracy and anger bias of children's faces. *Emotion* **2022**, *22*, 403. [CrossRef]
23. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*; First Trade Paperback Edition; PublicAffairs: New York, NY, USA, 2020.
24. Sajjad, M.; Nasir, M.; Ullah, F.U.M.; Muhammad, K.; Sangaiah, A.K.; Baik, S.W. Raspberry Pi assisted facial expression recognition framework for smart security in law-enforcement services. *Inf. Sci.* **2019**, *479*, 416–431. [CrossRef]
25. Laufs, J.; Borrion, H.; Bradford, B. Security and the smart city: A systematic review. *Sustain. Cities Soc.* **2020**, *55*, 102023. [CrossRef]
26. Rhue, L.A. Racial Influence on Automated Perceptions of Emotions. *CJRN Race Ethn.* 2018. Available online: https://racismandtechnology.center/wp-content/uploads/racial-influence-on-automated-perceptions-of-emotions.pdf (accessed on 26 August 2024).
27. Gleason, M. Privacy Groups Urge Zoom to Abandon Emotion AI Research. 2022. Available online: https://www.techtarget.com/searchunifiedcommunications/news/252518128/Privacy-groups-urge-Zoom-to-abandon-emotion-AI-research (accessed on 6 September 2024).
28. Stark, L.; Hoey, J. The ethics of emotion in Artificial Intelligence systems. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 3–10 March 2021; FAccT '21, pp. 782–793. [CrossRef]
29. Hernandez, J.; Lovejoy, J.; McDuff, D.; Suh, J.; O'Brien, T.; Sethumadhavan, A.; Greene, G.; Picard, R.; Czerwinski, M. Guidelines for Assessing and Minimizing Risks of Emotion Recognition Applications. In Proceedings of the 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), Nara, Japan, 28 September–1 October 2021; pp. 1–8. [CrossRef]
30. Fernández-Dols, J.M.; Russell, J.A. *The Science of Facial Expression*; Oxford Series in Social Cognition and Social Neuroscience; Oxford University Press: Oxford, UK, 2017.
31. Dixon, T. *From Passions to Emotions: The Creation of a Secular Psychological Category*; Cambridge University Press: Cambridge, UK, 2003.

32. Lin, W.; Li, C. Review of studies on emotion recognition and judgment based on physiological signals. *Appl. Sci.* **2023**, *13*, 2573. [CrossRef]
33. Roshdy, A.; Karar, A.; Kork, S.A.; Beyrouthy, T.; Nait-ali, A. Advancements in EEG Emotion Recognition: Leveraging multi-modal database integration. *Appl. Sci.* **2024**, *14*, 2487. [CrossRef]
34. James, W. What is an emotion? *Mind* **1884**, *9*, 188–205. [CrossRef]
35. Plutchik, R. *The Psychology and Biology of Emotion*; HarperCollins College Publishers: New York, NY, USA, 1994.
36. Plutchik, R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *Am. Sci.* **2001**, *89*, 344–350. [CrossRef]
37. Kleinginna, P.R.; Kleinginna, A.M. A categorized list of emotion definitions, with suggestions for a consensual definition. *Motiv. Emot.* **1981**, *5*, 345–379. [CrossRef]
38. Skinner, B.F. *Science and Human Behavior*; Macmillan: New York, NY, USA, 1953.
39. Oatley, K.; Johnson-Laird, P.N. Cognitive approaches to emotions. *Trends Cogn. Sci.* **2014**, *18*, 134–140. [CrossRef] [PubMed]
40. Radò, S. *Adaptational Psychodynamics: Motivation and Control*; Science House: New York, NY, USA, 1969.
41. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2019**, *10*, 18–31. [CrossRef]
42. Tomkins, S.S. *Affect Imagery Consciousness: The Complete Edition*; Springer Publisher: Berlin/Heidelberg, Germany, 2008.
43. Barrett, L.F. Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cogn. Emot.* **1998**, *12*, 579–599. [CrossRef]
44. Wundt, W.M. *An Introduction to Psychology*; G. Allen, Limited: London, UK, 1912.
45. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [CrossRef]
46. Russell, J.A.; Mehrabian, A. Evidence for a three-factor theory of emotions. *J. Res. Personal.* **1977**, *11*, 273–294. [CrossRef]
47. Mehrabian, A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Curr. Psychol.* **1996**, *14*, 261–292. [CrossRef]
48. Ekman, P.; Rosenberg, E.L. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*; Oxford University Press: Oxford, UK, 1997.
49. The European Parliament and the Council of the European Union. Artificial Intelligence Act. 2024. Available online: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf (accessed on 7 September 2024).
50. Stahl, B.C. Ethical Issues of AI. In *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*; Springer International Publishing: Cham, Switzerland, 2021; pp. 35–53. [CrossRef]
51. Cabitza, F.; Campagner, A.; Albano, D.; Aliprandi, A.; Bruno, A.; Chianca, V.; Corazza, A.; Di Pietto, F.; Gambino, A.; Gitto, S.; et al. The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability. *Appl. Sci.* **2020**, *10*, 4014. [CrossRef]
52. Barrett, L.F.; Westlin, C. Chapter 2—Navigating the science of emotion. In *Emotion Measurement*; Meiselman, H.L., Ed.; Woodhead Publishing: Sawston, UK, 2021; pp. 39–84. [CrossRef]
53. Zepf, S.; Hernandez, J.; Schmitt, A.; Minker, W.; Picard, R.W. Driver emotion recognition for intelligent vehicles: A survey. *ACM Comput. Surv.* **2020**, *53*, 1–30. [CrossRef]
54. Awatramani, J.; Hasteer, N. Facial expression recognition using deep learning for children with autism spectrum disorder. In Proceedings of the 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 30–31 October 2020; pp. 35–39.
55. Ullah, R.; Hayat, H.; Siddiqui, A.A.; Siddiqui, U.A.; Khan, J.; Ullah, F.; Hassan, S.; Hasan, L.; Albattah, W.; Islam, M.; et al. A real-time framework for human face detection and recognition in CCTV images. *Math. Probl. Eng.* **2022**, *2022*. [CrossRef]
56. Vardarlier, P.; Zafer, C. Use of Artificial Intelligence as business strategy in recruitment process and social perspective. In *Digital Business Strategies in Blockchain Ecosystems: Transformational Design and Future of Global Business*; Springer: Cham, Switzerland, 2020; pp. 355–373.
57. Chowdary, M.K.; Nguyen, T.N.; Hemanth, D.J. Deep Learning-based facial emotion recognition for human–computer interaction applications. *Neural Comput. Appl.* **2023**, *35*, 23311–23328. [CrossRef]
58. Huang, C.W.; Wu, B.C.; Nguyen, P.A.; Wang, H.H.; Kao, C.C.; Lee, P.C.; Rahmanti, A.R.; Hsu, J.C.; Yang, H.C.; Li, Y.C.J. Emotion recognition in doctor-patient interactions from real-world clinical video database: Initial development of artificial empathy. *Comput. Methods Programs Biomed.* **2023**, *233*, 107480. [CrossRef] [PubMed]
59. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
60. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
61. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
62. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional Neural Networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
63. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]

64. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
65. Pavez, R.; Diaz, J.; Arango-Lopez, J.; Ahumada, D.; Mendez-Sandoval, C.; Moreira, F. Emo-mirror: A proposal to support emotion recognition in children with autism spectrum disorders. *Neural Comput. Appl.* **2023**, *35*, 7913–7924. [CrossRef] [PubMed]
66. Silva, V.; Soares, F.; Esteves, J.S.; Santos, C.P.; Pereira, A.P. Fostering emotion recognition in children with autism spectrum disorder. *Multimodal Technol. Interact.* **2021**, *5*, 57. [CrossRef]
67. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three Machine Learning contests. In Proceedings of the Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Republic of Korea, 3–7 November 2013; Proceedings, Part III 20; Springer: Berlin/Heidelberg, Germany, 2013; pp. 117–124.
68. McStay, A. *Emotional AI: The Rise of Empathic Media*; Sage Publications Ltd.: Melbourne, VIC, Australia, 2018. [CrossRef]
69. Katirai, A. Ethical considerations in emotion recognition technologies: A review of the literature. *AI Ethics* **2023**, 1–22. [CrossRef]
70. Podoletz, L. We have to talk about emotional AI and crime. *AI Soc.* **2023**, *38*, 1067–1082. [CrossRef]
71. Spiroiu, F. The impact of beliefs concerning deception on perceptions of nonverbal Behavior: Implications for neuro-linguistic programming-based lie detection. *J. Police Crim. Psychol.* **2018**, *33*, 244–256. [CrossRef]
72. Finlay, A. *Global Information Society Watch 2019: Artificial Intelligence: Human Rights, Social Justice and Development*; Association for Progressive Communications (APC): Johannesburg, South Africa, 2019.
73. Qiang, X. President XI's surveillance state. *J. Democr.* **2019**, *30*, 53. [CrossRef]
74. Watch, H.R. China's Algorithms of Repression: Reverse Engineering a Xinjiang Police Mass Surveillance App. 2019. Available online: https://www.hrw.org/report/2019/05/01/chinas-algorithms-repression/reverse-engineering-xinjiang-police-mass (accessed on 28 May 2024).
75. Luca Zorloni. iBorderCtrl: La "Macchina Della Verità"' che l'Europa Userà ai Confini. 2023. Available online: https://www.wired.it/article/iborderctrl-macchina-verita-europa/ (accessed on 8 September 2024).
76. Carrer, L. Storia di un Ordinario Flop del Riconoscimento Facciale. 2024. Available online: https://www.wired.it/article/riconoscimento-facciale-fallimento-arresto-stadio/ (accessed on 5 July 2024).
77. Landowska, A. Uncertainty in emotion recognition. *J. Inf. Commun. Ethics Soc.* **2019**, *17*, 273–291. [CrossRef]
78. Canal, F.Z.; Müller, T.R.; Matias, J.C.; Scotton, G.G.; de Sa Junior, A.R.; Pozzebon, E.; Sobieranski, A.C. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Inf. Sci.* **2022**, *582*, 593–617. [CrossRef]
79. Ko, B.C. A brief review of facial emotion recognition based on visual information. *Sensors* **2018**, *18*, 401. [CrossRef] [PubMed]
80. Naga, P.; Marri, S.D.; Borreo, R. Facial emotion recognition methods, datasets and technologies: A literature survey. *Mater. Today Proc.* **2023**, *80*, 2824–2828. [CrossRef]
81. Mohanta, S.R.; Veer, K. Trends and challenges of image analysis in facial emotion recognition: A review. *Netw. Model. Anal. Health Inform. Bioinform.* **2022**, *11*, 35. [CrossRef]
82. Jones, M.; Viola, P. *Fast Multi-View Face Detection*; Mitsubishi Electric Research Lab TR-20003-96: Cambridge, MA, USA, 2003; Volume 3, p. 2.
83. Soo, S. *Object Detection Using Haar-Cascade Classifier*; Institute of Computer Science, University of Tartu: Tartu, Estonia, 2014; Volume 2, pp. 1–12.
84. Kumar, K.S.; Prasad, S.; Semwal, V.B.; Tripathi, R.C. Real time face recognition using AdaBoost improved fast PCA algorithm. *Int. J. Artif. Intell. Appl.* **2011**, *2*, 45–58. [CrossRef]
85. Rajesh, K.; Naveenkumar, M. A robust method for face recognition and face emotion detection system using support vector machines. In Proceedings of the 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), Mysuru, India, 9–10 December 2016; pp. 1–5.
86. Wang, Y.; Li, Y.; Song, Y.; Rong, X. Facial expression recognition based on random forest and convolutional Neural Network. *Information* **2019**, *10*, 375. [CrossRef]
87. Li, X.; Ji, Q. Active affective state detection and user assistance with dynamic Bayesian Networks. *IEEE Trans. Syst. Man-Cybern.-Part Syst. Humans* **2004**, *35*, 93–105. [CrossRef]
88. Mollahosseini, A.; Chan, D.; Mahoor, M.H. Going Deeper in facial expression recognition using Deep Neural Networks. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–10.
89. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1195–1215. [CrossRef]
90. Parkhi, O.; Vedaldi, A.; Zisserman, A. Deep face recognition. In Proceedings of the British Machine Vision Conference 2015, British Machine Vision Association, Swansea, UK, 7–10 September 2015.
91. Cabitza, F.; Ciucci, D.; Rasoini, R. A giant with feet of clay: On the validity of the data that feed Machine Learning in medicine. In *Proceedings of the Organizing for the Digital World*; Cabitza, F., Batini, C., Magni, M., Eds.; Springer: Cham, Switzerland, 2019; pp. 121–136.
92. Cabitza, F.; Campagner, A.; Sconfienza, L.M. As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 1–21. [CrossRef] [PubMed]
93. Li, S.; Deng, W. A deeper look at facial expression dataset bias. *IEEE Trans. Affect. Comput.* **2020**, *13*, 881–893. [CrossRef]

94. Yang, T.; Yang, Z.; Xu, G.; Gao, D.; Zhang, Z.; Wang, H.; Liu, S.; Han, L.; Zhu, Z.; Tian, Y.; et al. Tsinghua facial expression database—A database of facial expressions in Chinese young and older women and men: Development and validation. *PLoS ONE* **2020**, *15*, e0231304. [CrossRef] [PubMed]
95. Dalrymple, K.A.; Gomez, J.; Duchaine, B. The Dartmouth Database of Children's Faces: Acquisition and validation of a new face stimulus set. *PLoS ONE* **2013**, *8*, e79131. [CrossRef] [PubMed]
96. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
97. LoBue, V.; Thrasher, C. The Child Affective Facial Expression (CAFE) set: Validity and reliability from untrained adults. *Front. Psychol.* **2015**, *5*, 127200. [CrossRef] [PubMed]
98. Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Chen, F.; Wang, X. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Trans. Multimed.* **2010**, *12*, 682–691. [CrossRef]
99. Meuwissen, A.S.; Anderson, J.E.; Zelazo, P.D. The creation and validation of the developmental emotional faces stimulus set. *Behav. Res. Methods* **2017**, *49*, 960–966. [CrossRef]
100. Mavadati, S.M.; Mahoor, M.H.; Bartlett, K.; Trinh, P.; Cohn, J.F. DISFA: A Spontaneous Facial Action Intensity Database. *IEEE Trans. Affect. Comput.* **2013**, *4*, 151–160. [CrossRef]
101. Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D.H.; Hawk, S.T.; Van Knippenberg, A. Presentation and validation of the Radboud Faces Database. *Cogn. Emot.* **2010**, *24*, 1377–1388. [CrossRef]
102. Vemulapalli, R.; Agarwala, A. A compact embedding for facial expression similarity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5683–5692.
103. Benitez-Quiroz, C.F.; Srinivasan, R.; Feng, Q.; Wang, Y.; Martinez, A.M. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv* **2017**, arXiv:1703.01210.
104. Kosti, R.; Alvarez, J.M.; Recasens, A.; Lapedriza, A. Emotion Recognition in Context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
105. Barros, P.; Churamani, N.; Lakomkin, E.; Siqueira, H.; Sutherland, A.; Wermter, S. The OMG-emotion behavior dataset. In Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–7.
106. Nojavanasghari, B.; Baltrušaitis, T.; Hughes, C.E.; Morency, L.P. Emoreact: A multimodal approach and dataset for recognizing emotional responses in children. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 137–144.
107. Zafeiriou, S.; Kollias, D.; Nicolaou, M.A.; Papaioannou, A.; Zhao, G.; Kotsia, I. Aff-Wild: Valence and arousal 'in-the-wild' challenge. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1980–1987. [CrossRef]
108. Kollias, D.; Schulc, A.; Hajiyev, E.; Zafeiriou, S. Analysing affective behavior in the first ABAW 2020 competition. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 637–643. [CrossRef]
109. Gendron, M.; Barrett, L.F. Facing the past: A history of the face in psychological research on emotion perception. In *The Science of Sacial Expression*; Oxford Series in Social Cognition and Social Neuroscience; Oxford University Press: New York, NY, USA, 2017; pp. 15–36.
110. McStay, A.; Pavliscak, P. Emotional Artificial Intelligence: Guidelines for Ethical Use. 2019. Available online: https://emotionalai.org/outputs (accessed on 7 August 2024).
111. Crawford, K. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*; Yale University Press: New Haven, CT, USA, 2021.
112. Crawford, K. Time to regulate AI that interprets human emotions. *Nature* **2021**, *592*, 167. [CrossRef]
113. Keltner, Dacher and Ekman, Paul. The Science of "Inside Out". 2015. Available online: https://www.paulekman.com/blog/the-science-of-inside-out/ (accessed on 7 September 2024).
114. Matsumoto, D. American-Japanese cultural differences in the recognition of universal facial expressions. *J. Cross-Cult. Psychol.* **1992**, *23*, 72–84. [CrossRef]
115. Matsumoto, D.; Yoo, S.H.; Nakagawa, S. Culture, emotion regulation, and adjustment. *J. Personal. Soc. Psychol.* **2008**, *94*, 925. [CrossRef] [PubMed]
116. Matsumoto, D. Cultural similarities and differences in display rules. *Motiv. Emot.* **1990**, *14*, 195–214. [CrossRef]
117. Butler, E.A.; Lee, T.L.; Gross, J.J. Emotion regulation and culture: Are the social consequences of emotion suppression culture-specific? *Emotion* **2007**, *7*, 30. [CrossRef] [PubMed]
118. Barrett, L.F. The theory of constructed emotion: An active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* **2016**, *12*, 1–23. [CrossRef] [PubMed]
119. Floridi, L. *Etica dell'Intelligenza Artificiale: Sviluppi, Opportunità, Sfide*; Raffaello Cortina Editore: Milano, Italy, 2022.
120. Booth, B.M.; Hickman, L.; Subburaj, S.K.; Tay, L.; Woo, S.E.; D'Mello, S.K. Integrating psychometrics and computing perspectives on bias and fairness in Affective Computing: A case study of automated video interviews. *IEEE Signal Process. Mag.* **2021**, *38*, 84–95. [CrossRef]
121. Reyes, B.N.; Segal, S.C.; Moulson, M.C. An investigation of the effect of race-based social categorization on adults' recognition of emotion. *PLoS ONE* **2018**, *13*, e0192418. [CrossRef]

122. Hutchings, P.B.; Haddock, G. Look Black in anger: The role of implicit prejudice in the categorization and perceived emotional intensity of racially ambiguous faces. *J. Exp. Soc. Psychol.* **2008**, *44*, 1418–1420. [CrossRef]
123. Kim, E.; Bryant, D.; Srikanth, D.; Howard, A. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtually, 19–21 May 2021; pp. 638–644.
124. Xu, T.; White, J.; Kalkan, S.; Gunes, H. Investigating bias and fairness in facial expression recognition. In Proceedings of the Computer Vision–ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; Proceedings, Part VI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 506–523.
125. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; pp. 77–91.
126. Drozdowski, P.; Rathgeb, C.; Dantcheva, A.; Damer, N.; Busch, C. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Trans. Technol. Soc.* **2020**, *1*, 89–103. [CrossRef]
127. Stark, L. The emotional context of information privacy. *Inf. Soc.* **2016**, *32*, 14–27. [CrossRef]
128. McStay, A. Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. *Big Data Soc.* **2020**, *7*, 205395172090438. [CrossRef]
129. Sánchez-Monedero, J.; Dencik, L. The politics of deceptive borders: 'Biomarkers of deceit' and the case of iBorderCtrl. *Inf. Commun. Soc.* **2022**, *25*, 413–430. [CrossRef]
130. Kalantarian, H.; Jedoui, K.; Washington, P.; Tariq, Q.; Dunlap, K.; Schwartz, J.; Wall, D.P. Labeling images with facial emotion and the potential for pediatric healthcare. *Artif. Intell. Med.* **2019**, *98*, 77–86. [CrossRef] [PubMed]
131. Nagy, J. Autism and the making of emotion AI: Disability as resource for surveillance capitalism. *New Media Soc.* **2024**, *26*, 14614448221109550. [CrossRef]
132. European Parliament. EU AI Act: First Regulation on Artificial Intelligence. 2023. Available online: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed on 1 September 2024).
133. NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023. Available online: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf (accessed on 7 August 2024).
134. European Parliament. Amendments Adopted by the European Parliament on 14 June 2023 on the Proposal for a Regulation of the European Parliament and of the Council on Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. 2023. Available online: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf (accessed on 6 August 2024).
135. Council, N.R. *The Polygraph and Lie Detection*; The National Academies Press: Washington, DC, USA, 2003. [CrossRef]
136. Hayes, A.F.; Krippendorff, K. Answering the call for a standard reliability measure for coding data. *Commun. Methods Meas.* **2007**, *1*, 77–89. [CrossRef]
137. Chen, Y.; Joo, J. Understanding and Mitigating Annotation Bias in Facial Expression Recognition. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 14960–14971. [CrossRef]
138. American Psychological Association. The Truth about Lie Detectors (Aka Polygraph Tests). 2004. Available online: https://www.apa.org/topics/cognitive-neuroscience/polygraph (accessed on 6 August 2024).
139. Leahu, L.; Schwenk, S.; Sengers, P. Subjective objectivity: Negotiating emotional meaning. In Proceedings of the 7th ACM Conference on Designing Interactive Systems, Cape Town, South Africa, 25–27 February 2008; pp. 425–434.
140. Faigman, D.L.; Fienberg, S.E.; Stern, P.C. The limits of the polygraph. *Issues Sci. Technol.* **2003**, *20*, 40–46.
141. Nortje, A.; Tredoux, C. How good are we at detecting deception? A review of current techniques and theories. *S. Afr. J. Psychol.* **2019**, *49*, 491–504. [CrossRef]
142. U.S. United States v. Scheffer. Opinions and Dissents, Supreme Court. 1998. Available online: https://supreme.justia.com/cases/federal/us/523/303/ (accessed on 6 August 2024).
143. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*; Sage Publications Sage: Thousand Oaks, CA, USA, 2018. [CrossRef]
144. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Trans. Image Process.* **2018**, *28*, 2439–2450. [CrossRef] [PubMed]
145. The European Parliament and the Council of the European Union. General Data Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC. 2016. Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed on 26 August 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.