Guidelines for Assessing and Minimizing Risks of Emotion Recognition Applications

Javier Hernandez[+*1], Josh Lovejoy[*1], Daniel McDuff[1], Jina Suh[1], Tim O'Brien[1], Arathi Sethumadhavan[1], Gretchen Greene[2], Rosalind Picard[3], and Mary Czerwinski[1]

[1]Microsoft Research, Microsoft, Redmond, USA
[2]Harvard Kennedy School, Harvard University, Cambridge, USA
[3]Media Lab, Massachusetts Institute of Technology, Cambridge, USA

***Abstract*—Society has witnessed a rapid increase in the adoption of commercial uses of emotion recognition. Tools that were traditionally used by domain experts are now being used by individuals who are often unaware of the technology's limitations and may use them in potentially harmful settings. The change in scale and agency, paired with gaps in regulation, urge the research community to rethink how we design, position, implement and ultimately deploy emotion recognition to anticipate and minimize potential risks. To help understand the current ecosystem of applied emotion recognition, this work provides an overview of some of the most frequent commercial applications and identifies some of the potential sources of harm. Informed by these, we then propose 12 guidelines for systematically assessing and reducing the risks presented by emotion recognition applications. These guidelines can help identify potential misuses and inform future deployments of emotion recognition.**

***Index Terms*—ethics, guidelines, risk, emotion recognition**

I. Introduction

The field of Affective Computing has experienced significant growth since its original inception in 1997 with the book by the same name [1]. While most of the early research was performed in small, controlled laboratory experiments with custom-made and expensive sensing technologies [2], [3], we are now seeing an increasing number of real-life deployments that leverage low-cost, mass-produced sensors [4], [5]. This shift has been accompanied by a significant democratization and popularization of emotion sensing technologies, APIs and services that can extract emotional information from images [6], [7], voice [8], text [9] and vital signs [10]. Unfortunately, affective computing, and especially the area of emotion recognition, has been used in unexpected and potentially harmful settings by practitioners without sufficient understanding of its limitations or assumptions. We believe that the lowered barriers to entry for using emotion recognition, the increased agency of non-expert practitioners, and the gaps in regulating the usage of this technology create a demand for the research community to question how we design, position, implement and ultimately deploy these tools.

How should emotion recognition label each of the Serena Williams' photos in Figure 1? An algorithm that only sees the image on the left is likely to label it as "anger", as it may detect the mouth stretch and nose wrinkle, but if that same algorithm were to see the image on the right, it is likely to label it as "excitement" instead, as it may recognize that Serena is celebrating a victory with a fist pump. While one could potentially argue that the second prediction may be more accurate as it has access to more contextual information, we could all potentially agree that none of the algorithms can really know how Serena Williams was feeling at that precise moment without asking her. Even if we are able to ask her, we know she may still be biased by the way we framed our question [12], potential recall biases [13], and most likely some misattribution of arousal [14]. While these observations may seem obvious to the general affective computing researcher, we have observed that the terminology and the models used to represent emotion recognition technology can be confusing and that new practitioners might overestimate the predictive power of AI [15]. For instance, commercial facial expression analysis often adopts Ekman's theory of basic emotions [16] that suggests that certain expressions may be universally indicative of emotional states and, consequently, would oversimply the subtleties of the previous example. Similar algorithms are being applied in various settings, such as mental health support and job recruitment, without the appropriate communication of system capabilities, limitations and contextual understanding. This can lead to significant user harm, such as mental health deterioration or career opportunity loss, respectively.

Fig. 1. How should emotion recognition label each of these images? Example adapted from Barrett et al. [11] illustrating the importance of context.

[+]Corresponding author: javierh@microsoft.com
[*]Both authors contributed equally to this work

This is just one example that demonstrates that the positioning of emotion recognition—both its capabilities and its limitations—plays a critical role in its potential deployment and commercial adoption. Without clear guidelines, there is a real risk of the technology causing harm. As a first step towards mitigating harm through the use of emotion recognition technologies, our work outlines commercial and real-life applications of emotion recognition and their potential sources of challenges and user harm (Section III). We then propose a set of ethical application guidelines grounded on four main regulating factors of risk to systematically assess risks and minimize harm (Sections IV,V). Finally, our work contributes to the growing number of efforts investigating the role of AI in our society in order to motivate a more principled and responsible use of emotion recognition technologies.

II. Related Work

The ethics of affective computing has been an active research topic since the inception of the field [1], [17]–[23]. Many of these studies often consider the challenges associated with the deployment of sentient machines that can effectively read and simulate emotions (e.g., intimate robots [24], [25], mindreading [26]). However, current commercial applications of emotion recognition are still far from these scenarios due to well-known research challenges such as the idiosyncratic nature of emotion perception, the need for better emotion theory, and the difficulty of creating truly representative models [17]. Despite knowing these limitations, Stark and Hoey [27] recently argued that current applications of emotion recognition are typically designed without awareness of the ethical complexities involved in emotion recognition technology nor with an understanding of the paucity of a globally accepted, unified theory of emotion. Similarly, several independent reports have examined the potential challenges of emotion recognition technologies and have made a call for action, such as urging to revising the ACM code of ethics [28], regulating the technology [29], [30], and having more conversations across fields [31]. Despite the increasing debate and shared concerns around user harm, however, there is still a need for consensus on concrete guidelines that help anticipate and address the challenges associated with emotion recognition.

To help address similar challenges across other AI domains, researchers have devised important guidelines that have facilitated the understanding and prevention of AI risks. For instance, Chancellor et al. [32] reviewed some of the risks associated with mental health research from social media and identified 14 ethical challenges as well as recommendations for researchers such as carrying out participatory algorithmic design. In Amershi et al. [33], the authors synthesized over 20 years of guidance around the design of human-AI interaction and provided a set of 18 AI usability guidelines to help design more useful and acceptable AI-based products and prototypes. In a separate effort, Batliner et al. [34] reviewed some exemplary use cases of computational paralinguistics and provided 6 rules for good practice such as facilitating model interpretability and result comparison. Finally, Stark and Hoey [27] used case studies around emotionally intelligent cognitive assistance technologies and agents in social networks to consider ethical practices in emotion recognition. In particular, they reviewed an independent checklist for ethical emotion recognition designed by McStay and Pavliscak [35] that provides 15 guidelines distributed across personal, relationship, and societal implications. While it is unclear exactly how the authors derived and evaluated their guidelines, we see them as complementary to the ones proposed in this work.

III. Applications of Emotion Recognition

Emotion recognition technology can be a beneficial tool, but it can also introduce societal challenges when used in unexpected or harmful settings. To contextualize how and where these challenges may occur, we first provide an overview of recent commercial applications of the technology, outline some main sources of challenges for emotion recognition, and then characterize potential harms originating from these challenges.

A. Application Areas

To help understand how emotion recognition is currently being used in real-life settings, we compiled a non-exhaustive list of some of the most frequently explored applications. To identify these applications, we searched for "commercial emotion recognition" and "applications of emotion recognition" using popular search engines, and reviewed the top 20 results to assess whether they described 1) the use of emotion recognition, 2) in a real-life setting, and 3) in a commercial setting. If the results referenced other articles, we added them to the reviewing list. While we did not impose any limitation in terms of language or countries, it is important to note that our keyword selection could have introduced a bias towards English-speaking countries (e.g., USA, UK). In addition, we observed that more highly ranked articles tended to be more critical which may have also biased the sampling towards more sensitive applications. Table I shows the resulting list of applications which can be grouped into the following areas:

Market Research. One early commercial application of emotion recognition is in marketing and consumer research. The goal is to help marketers understand consumers in order to deliver more relevant and personalized products [36], [37]. Similarly, textual analysis has been used to analyze the affect or sentiment of product reviews [9]. Leveraging this type of information, content providers like Netflix [38] and Spotify [39] provide emotional filters as part of their recommendations.

Human Resources. Several companies have started using emotion recognition to gather insights about current and prospective employees. In hiring, this technology has been marketed as a way of augmenting the screening of candidates to improve efficiency [40], [41]. In the workplace, the technology has been sold as a means of helping gauge the happiness of employees in relation to different events [42].

Transportation. Given that driving can often elicit negative emotional states, emotion recognition is increasingly being considered in the context of transportation. The transportation sector is now beginning to adopt this for consumer-facing

TABLE I
COMMERCIAL EMOTION RECOGNITION APPLICATIONS

Market sector	Applications
Market Research	Video response analysis [36], [37], sentiment of reviews [9], emotion filters [38], [39]
Human Resources	Candidate recruitment [40], [41], employee monitoring [42]
Transportation	Safety [43], entertainment [44], [45]
Defense and security	Deception detection [46], suspicious behavior [47], crowd analytics [48], [49],
Education	Student engagement [50], online proctoring [51], skill training [52], [53]
Mental health	Stress [55], [56], loneliness [60], depression [57], [59], suicide [58]

features such as accident prevention [43] and driving experience personalization [44], [45].

Defense and Security. The identification and prevention of potential security threats has long been of interest to governmental agencies. The possibility of detecting signs of deception during interviews from physiological and facial responses has been explored [46], and it has been suggested such tools might be deployed in airports [47]. Beyond this, it has been proposed that large scale measurement could be used to understand the potential state of the enemy and help inform decision making in warfare [48], [49].

Education. A long-standing and rapidly increasing domain of application for emotion recognition technology is in educational settings, with the intent to improve learning or the learning experience. For instance, teachers might improve learning experiences with insights from assessing the "engagement" of learners [50]. More recently, emotion recognition technologies have been marketed for online proctoring during remote exams sparking some concerns [51]. Emotion recognition has also been used to support the delivery of custom training and coaching, such as offering tips about presentation style and identifying characteristics about how an audience is responding [52]–[54].

Mental Health. Emotion recognition technologies are being used to help provide more frequent and personalized support for people in both clinical and non-clinical settings. Many of these technologies are used with the intent to help track and manage different states such as stress [55], [56], depression [57], suicidal ideation [58] and pre- and post-natal depression [59]. This trend has extended to robotics, where embodied agents are used to help reduce loneliness, especially for the elderly [60].

Given the steep rise in the number of start-ups in emotion recognition, it is not possible to exhaustively cover all of the applications here but there have also been relevant examples across domains such as entertainment [61], [62], customer service [63], [64], retail tracking [65], [66], urban space design [67], and interface control [68].

B. Sources of Challenges

As emotion recognition technologies become more prevalent, it is critical to identify ways in which they could lead to harm. The following section aims to highlight challenges that are especially acute to emotion recognition.

The theory of human emotions is evolving. In certain areas of AI, such as object recognition and classification, concepts and categories that delineate decision boundaries can be well-defined and agreed upon by humans in most cases (e.g., people would recognize and agree upon the label of a car given a picture of a car). In contrast, emotion recognition relies heavily on psychological theories to inform measurement, representation, and modeling. However, emotion theory continues to evolve and be hotly debated [69], [70]. Furthermore, there are cultural, geographical and individual differences that influence how we may perceive, experience, and express emotions [4], [71], [72] and these are moderated by context [11], [70].

Human emotions are difficult to describe and label. Emotions are internal states, but they may be associated with external behaviors (e.g., facial expressions, non-verbal gestures) and measurable physiological responses (e.g., perspiration, heart rate). But to provide a symbolic label for an emotion (e.g., a word or description) it needs to be self-reported by the subject (e.g., "I am happy") or inferred by an observer (e.g., "You look stressed."). Differences in individual styles or cultural norms for emotional expressiveness and suppression can lead to an obvious mismatch between felt emotions, expressed emotions, and perceived emotions [71]–[73]. Unfortunately, this incongruency has often been neglected when creating data sets for training emotion recognition technologies and labeling is often performed based only on a perceiver's, or perceivers', interpretation of the externalized expressions. Some emotion recognition systems have been trained on subjective self-reports. However, self-reports are vulnerable to a wide variety of confounds that can decrease their quality and reproducibility. Examples of potential confounds include alexithymia, recall bias, power asymmetries between individuals, or the framing of the questions [12]–[14].

A lack of representative and generalizable data. The complexities of emotion, along with the variability in how they are experienced, leads to large variability. To control for factors leading to these variabilities, emotion recognition models in the research context are usually developed and evaluated within well-defined experimental contexts (e.g., sedentary positions, frontal faces). However, these models are often deployed in contexts that may vary significantly outside the distribution of the training conditions (e.g., high activity level, lateral faces, an older population). Even though the models may perform well against one particular test set, they may not perform well when deployed in real-life settings. For example, some facial analyses have shown poor generalization performance on data sets containing darker skin tones [74]–[77].

Oversimplified language is used to communicate system capabilities. The terminology used to describe emotion recognition systems often elicits insufficient, and sometimes inaccurate,

understanding of their capabilities and obscures their technical limitations. For instance, most emotion sensing APIs fail to disclose information about the data sets that were used to train and validate their methods. Similarly, many of the offerings do not differentiate between internal emotional states, externalized expressions, self-reported states, and perceived emotions, and fail to describe their limitations. Further, what researchers use to carefully describe what emotion recognition systems do gets oversimplified to the general public, e.g., "recognizing patterns and mapping them to likely labels as perceived by humans in a specific training context" as "reading inner feelings," promoting confusion and technological missattributions.

There is a blurred boundary between what should be private and public. Emotions play a critical role in multiple facets of our lives, ranging from regulating external social interaction to the processing of our internal states. Although the former is an external and conscious "use" of emotions, the latter occurs in a way that is virtually invisible to external perceivers. However, current emotion recognition systems tend to blur the distinction between the two and do not appropriately account for the potentially unique set of expectations about what is and/or should be made public and what should remain private. Understanding users' expectations in the context of emotion sensing is especially difficult because there are large cultural and societal norms and differences [71], [73], [78].

C. Potential Harms

As commercial AI applications have grown, several efforts have focused on understanding and characterizing the types of harms they may create or incite. Among the different efforts, this work leverages Azure's types of harms [79] and identifies which harms are more prevalent in the context of emotion recognition. This taxonomy draws from prior efforts such as Value Sensitive Design [23] and was designed for AI applications in general.

Denial of consequential services. This type of harm appears when the use of AI can influence the access to services and opportunities by an individual. Considering the previous applications, this type of harm is more likely to appear in uses of emotion recognition that attempt to evaluate someone's fitness in the context of an application, such as social services, credit/loans, or job opportunities [40], [41]. For instance, people who are underrepresented in a data set (e.g., neuroatypical, minorities) will not be evaluated equally as those that are more represented [76], [77]. This may contribute to stereotype reinforcements that preserve the denial of consequential services [80].

Risk of injury. This type of harm appears when the use of AI can lead to physical or emotional damage to an individual. For instance, physical harm can be elicited when users overly rely on safety mechanisms or alerts, such as those built into cars to help prevent accidents [43]. Applications of emotion recognition may be more prone to emotional and psychological harm. For instance, users of emotional support apps may become vulnerable if they over-rely on AI systems rather than a trained mental health counselor. People watching TV or children playing with toys may be vulnerable to attention hijacking if emotion recognition is designed to sense and promote prolonged user interaction. Furthermore, content providers that leverage emotion recognition may incentivize the development and recommendation of content that is manipulative, polarizing and/or deceptive. In addition, people may experience reputation damage when emotion recognition technologies are used to publicly analyze their emotions (e.g., sentiment analysis of reviews [9], emotional analysis of debates).

Infringement on human rights. This type of harm appears when the use of AI can impact the privacy, freedom, and/or rights of users [81]. For instance, emotion recognition applied to lie detection [46], employee monitoring [42] or public transport [45] can cause interference with private life by capturing information that the user did not consent to share. Similarly, emotion recognition can be indiscriminately applied for predictive policing such as suspicious behavior detection [47] or cheating detection [51]. Emotion recognition can also prevent users from freely and fully developing themselves if they heavily rely on the outputs of imperfect technology to direct their approach to mental health self care. This harm may also happen if users experience a sense of forced conformity in order to align with a larger group distribution's actions, as is likely to happen when monitoring behavior in social scenarios. Finally, emotion recognition used in the context of content recommendation [38], [39] could lead to the marginalization of minority perspectives as those are less frequently represented.

IV. Ethical Application Guidelines

The previous section has identified relevant challenges of emotion recognition as well as potential harms they may elicit. Next, we propose a set of ethical guidelines (see the checklist in Table II) for emotion recognition applications that highlight four main factors that regulate risk–we will refer to these as the 4Cs: *communication*, *consent*, *calibration*, *contingency*. For each factor, we introduce three critical considerations that highlight some of the most important areas specific to emotion recognition applications.

A. Responsible Communication

System design should facilitate responsible communication with the users, which is critical in the context of emotion recognition due to the complexity and subtleties of emotions and emotion recognition tools. This communication can happen at different points during the interaction, such as when describing the general capabilities of emotion recognition technology and/or when providing predictions to the users. In particular, we identify the following key considerations:

G1. Predictions. Affective outputs are used and described in a manner that does not presume to represent the ground truth about a user's emotions and avoids making value judgments about others.

G2. Granularity. Descriptions about emotion recognition should use specific language that can be understood by users depending on their context, tasks, and objectives.

G3. Transparency. Information provided to the user regarding emotion recognition should promote awareness about the purpose of its use, limitations, and intended/supported scenarios (e.g., demographics of target users, activities being performed).

B. Informed Consent

An informed consent functions as a tool for practitioners and end-users to evaluate system capabilities and topics such as data handling, privacy, benefits, and risks of using the system, as well as promoting freedom of choice. Consent is relevant in the context of emotion recognition because (a) each user may have unique preferences and expectations regarding the disclosure of emotional states and (b) there is a potential power asymmetry between the person providing the consent and the end user which may be perceived as coercive (e.g., manager and direct report, professor and student), (c) individuals may be unaware of the types of physiological and behavioral data can be sensed (e.g., that a camera can measure their heart rate [82]). Therefore, we identify the following guidelines to help provide informed consent:

G4. Opt-in. Users and bystanders are provided with the information necessary to knowingly and willfully opt-in before any measurements are attempted about their emotional states.

G5. Comprehension. The consent request promotes awareness and understanding (i.e., the user is able to describe what they are consenting to in their own words) about how the application is processing, sharing and storing the user's data, and how the results will be used.

G6. Freedom of choice. The consent provides the freedom of choice and able to decline the measurement of their emotions without losing access to consequential opportunities.

C. Contextual Calibration

Appropriately calibrating machine learning models so they can effectively work in real-life settings is very challenging and requires consideration throughout the model development lifecycle (e.g., validation, deployment, customization). This is important when designing emotion recognition systems as there are often large individual and cultural differences as well as relevant contextual factors. To help ensure appropriate calibration of the models, we identify the following considerations:

G7. Representativeness. Models should be trained on a data set that is representative of the target population distribution and results are broken down by relevant demographics.

G8. Variance. Models should account for varying factors, such as individual and demographic differences, that influence the emotional experience and manifestation of emotions.

G9. Customization. Users are empowered to provide feedback on performance and usefulness to facilitate further customization of the models or the system.

D. Comprehensive Contingency

AI systems applied in real-life will inevitably make mistakes that may contribute to harm. In the context of emotion recognition, the space of potential errors can be large due to the diversity of possible contexts, individual differences, and interpretations. How system errors are handled can impact the regulating factors listed here. To help anticipate and prevent these errors, we identify the following considerations:

G10. Deletion. Users are able to review and redact emotion measurements that are personally identifiable.

G11. Feedback. Users should have accessible and inclusive channels of communication that facilitate sharing their experience anonymously. These channels should clearly indicate who is accountable for outcomes affected by emotion measurements.

G12. Robustness. The system is able to detect out-of-distribution data characteristics that may impair the performance of the models.

TABLE II
4CS GUIDELINES FOR EMOTION RECOGNITION APPLICATIONS

Responsible Communication
G1. **Predictions** are not handled as ground truth
G2. System descriptions should be described with **granularity**
G3. Technology should be described with **transparency**
Informed Consent
G4. **Opt-in** is facilitated before measurements are performed
G5. Data handling is described to facilitate **comprehension**
G6. Consent facilitates **freedom of choice** without consequences
Contextual Calibration
G7. Training data is **representative** of real-life data
G8. Sources of **variance** are accounted by the models
G9. Users can **customize** the system by providing feedback
Comprehensive Contingency
G10. Personal data can be **deleted** by the user
G11. **Feedback** channels are provided to the users
G12. Shifts in data distribution are detected to ensure **robustness**

V. RISK MITIGATION

The above guidelines can be used to proactively address the risks of emotion recognition before deployment. However, addressing some of these may be challenging depending on the scope and context of the application. Here we provide recommendations that can help mitigate some of the risks.

Avoiding assessments. While making predictions is an essential part of emotion recognition algorithms, the practitioner designing the application may decide to avoid surfacing explicit labels to help minimize risk. For instance, a stress management system may use predicted stress levels as a signal to interact with a user or trigger an intervention of some sort, but the application may not expressly label the user's experience as "stressed." Instead, it may indicate that changes in physiology have been detected and let the user make further interpretations based on their unique contextual and personal information. This approach, also known as Affective Interaction [83], is relevant when addressing the guidelines associated with Communication (G1-3) and the user is entirely in control of the information. However, this approach can still lead to important privacy and discrimination risks if shared with others.

Private by default. While not unique to emotion recognition, data minimization is an essential building block for reducing risk. To avoid falling into the trap of potential infringement on

human rights or causing psychological distress, it is safer to presume that individuals consider their emotional states private. Consequently, we recommend starting from the default position that emotion recognition data is for personal consumption and reflection. If some data needs to be shared in order to return clear and obvious benefit to the user, it is recommended to do so in aggregated and/or differentially private forms that maintain plausible deniability and/or obfuscate identifiable details of individuals. This approach is relevant when addressing the guidelines associated with Consent (G4-6).

Emotion as a form of expression. While one of the unrealized ambitions of emotion recognition is to understand the internal experience of emotions, we believe that focusing measurement on externalized manifestations of emotional communication can help ameliorate risk as these tend to be more public in nature and therefore may be more consciously controllable by actors. In the context of driving, for instance, an automotive safety application may detect "closed eyes" rather than "drowsiness" from face images. In the context of media indexing, an algorithm may decide to label an image "downturned mouth" rather than "sad person." For the cases when focusing on externalized manifestations is just not possible, modifications of the labels such as "perceived sadness" can help better position some of the capabilities of the technology, such as the goal of reflecting annotators' perceptions in the context of image content analysis. This approach is relevant when addressing the guidelines associated with Communication (G1-3) as well as Calibration (G7-9).

Human computer collaboration. When using emotion recognition in real-life scenarios, it is essential to understand that a wide variety of unexpected human behaviors will be encountered, as will various environmental conditions, edge device characteristics, and more. This is especially important in consequential scenarios such as mental healthcare, job recruitment, or online proctoring. To help mitigate erroneous system outputs introduced by previously unseen behaviors, we encourage a human computer collaboration model in which predictions are provided to an expert operator (e.g., clinician, interviewer, proctor) with the expectation that they will have the time, support, and situational awareness to reconcile unexpected behaviors and technology limitations. These machine "teachers" would be prepared to be personally accountable for the eventual decisions and outcomes of the system. It is important to note, however, that experts may still be influenced by important factors such as automation bias. This approach is relevant when addressing the guidelines associated with Contingencies (G10-12), but would also help facilitate more appropriate opt-out strategies (G6) and avoid making value judgments (G1).

VI. Discussion

The adoption of emotion recognition is rapidly increasing. Applications vary from measuring consumer behavior to assisting in making important decisions, such as supporting the mental health of people, promoting driver safety, and filtering job candidates. As emotion recognition tools become more ubiquitous, there is potential for misuse and harm. This work reviews the current landscape of emotion recognition applications, as well as some of the most common types of harm, such as denial of consequential services, physical and emotional injury, and infringement on human rights. Some of the central challenges include that emotion recognition is based on theories that are still evolving, that the ground truth of emotions is not easily quantifiable, and that large individual and cultural differences exist in terms of emotional experience and expression [72]. Further, the language used to communicate about emotion recognition technologies often over-promises what it can be used accurately for, and rarely reflects the possibility of potential misuse and automation bias.

To help address some of the previous challenges, we identified four key regulating factors of tension and human risks. These include technology communication, user consent, model calibration, and contingency measures, and are further decomposed into three main guidelines for each factor. However, it is important to note that these guidelines only tackle especially sensitive areas that are relevant for emotion recognition, and they do not represent an exhaustive list. Therefore, we recommend augmenting this list with traditional Institutional Review Board reviews, legal regulatory frameworks, and/or other guidelines that help capture sources of sensitivity in specific application areas. For instance, it is important that practitioners carefully evaluate the specific application area and assess whether the particular use of AI (emotion recognition in our case) is the right mechanism to achieve their intended results as well as ensure that the benefits for users outweigh the potential harms. This work also provides some recommendations to help effectively address some of the proposed guidelines, like adopting a human-computer collaboration model to appropriately interpret models' predictions and detect potential errors, or adopting a "private by default" position on data handling to ensure users' expectations are met, regardless of their demographics. There is also a wealth of methodology to help address specific guidelines, such as advancements in domain adaptation, to help account for sources of variance (G8), methods to assess the similarity between data sets (G7 and G12), and model transparency and interpretability (G1 and G3). As our understanding of the uses of emotion recognition continues to advance, we expect to see an increase in the number of regulating factors and guidelines, as well as more potential mitigating mechanisms.

The main goal of these guidelines is to offer a simple yet effective method to prevent or mitigate possible harms in the sensitive area of emotion recognition. It is important to note, however, that not all the guidelines may be applicable in every context and it is important to consider the potential for good on the end users [19], [22]. In addition, some of the specific guidelines may be difficult to address. Certain applications of emotion recognition, like automatically selecting relevant candidates or detecting cheating behavior, are currently being applied in settings where equitable power dynamics cannot be easily guaranteed (G6). In some cases, re-positioning the current strategy *could* help ensure that the guidance can be met. For instance, emotion recognition might be used to assist

interviewers to conduct more effective interviews, or to help proctors orient their attention better while overseeing test taking, respectively. When applying the guidelines, we also recommend addressing the four areas in tandem to help facilitate better control over the different factors. However, this is not always possible in real-life settings, due to the decentralization of the different factors. For instance, one of the main vehicles of emotion recognition democratization comes from companies that provide emotion sensing services, or APIs, so that other companies can develop their own end-user applications. This is particularly problematic as the same language and descriptions of the technology are often shared across a wide variety of applications, irrespective of their sensitivity and particular context. In addition, domain experts who develop the technology and are therefore more familiar with its capabilities and limitations, do not usually have direct control over how the emotion recognition technology will be used in practice. To help address the latter, it is critical to develop deeper engagement across service and solution providers, and encourage responsible communication that promotes awareness about the technology in particular, and emotions in general. We believe important research will need to systematically evaluate effective methods of conveying the unique complexities of emotions and the necessary technological guardrails to diagnose and prevent potential harm.

VII. Conclusion

Emotion recognition is increasingly being used in many commercial applications but can result in important harm to people and society. To help better guide future innovation in the space, this work surveyed applications of emotion recognition, articulated many potential sources of harm, and provided a set of guidelines and mitigation mechanisms to help detect and minimize risks in practice.

Acknowledgment

We would like to thank Jacquelyn Krones, Mira Lane, Steven Bowles, Ece Kamar, Mark Van Hollebeke, Kristen Laird, Michael Brent, and Hanna Wallach for sharing their insights on an earlier version of the guidelines, and Kate Crawford for reviewing the manuscript and providing helpful comments.

References

[1] R. W. Picard, *Affective Computing*. The MIT Press, 1997.

[2] J. Riseberg, J. Klein, R. Fernandez, and R. W. Picard, "Frustrating the user on purpose," in *CHI 98 Conference Summary on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery (ACM), apr 1998, pp. 227–228.

[3] J. Scheirer, R. Fernandez, and R. W. Picard, "Expression glasses: a wearable device for facial expression recognition," in *Extended Abstracts of Human Factors in Computing Systems*, 1999, pp. 262–263.

[4] D. McDuff and S. Glass, "Faded Smiles? A Largescale Observational Study of Smiling from Adolescence to Old Age," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 53–65, jan 2021.

[5] E. Smets, E. Rios Velazquez, G. Schiavone, I. Chakroun, E. D'Hondt, W. De Raedt, J. Cornelis, O. Janssens, S. Van Hoecke, S. Claes, I. Van Diest, and C. Van Hoof, "Large-scale wearable data reveal digital phenotypes for daily-life stress detection," *npj Digital Medicine*, vol. 1, no. 1, pp. 1–10, dec 2018.

[6] "Affectiva - Humanizing Technology: Affectiva." https://www.affectiva.com/

[7] "Perceived Emotion Recognition Using the Face API - Xamarin - Microsoft Docs." https://docs.microsoft.com/en-us/xamarin/xamarin-forms/data-cloud/azure-cognitive-services/emotion-recognition

[8] "openSMILE - audEERING." https://www.audeering.com/opensmile/

[9] "Brand24 - Media Monitoring Tool." https://brand24.com/

[10] "B-Secur - HeartKey." https://www.b-secur.com/heartkey/

[11] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in Emotion Perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, oct 2011.

[12] G. N. Yannakakis, R. Cowie, and C. Busso, "The Ordinal Nature of Emotions: An Emerging Approach," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 16–35, jan 2021.

[13] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment." *Annual Review of Clinical Psychology*, vol. 4, pp. 1–32, 2008.

[14] R. C. Sinclair, C. Hoffman, M. M. Mark, L. L. Martin, and T. L. Pickering, "Construct accessibility and the misattribution of arousal: Schachter and Singer Revisited," *Psychological Science*, vol. 5, no. 1, 1994.

[15] M. L. Cummings, "Automation bias in intelligent time critical decision support systems," in *Collection of Technical Papers - AIAA 1st Intelligent Systems Technical Conference*, vol. 2, 2004, pp. 557–562.

[16] R. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[17] R. W. Picard, "Affective computing: Challenges," *International Journal of Human Computer Studies*, vol. 59, no. 1-2, pp. 55–64, jul 2003.

[18] S. B. Daily, M. T. James, D. Cherry, J. J. Porter, S. S. Darnell, J. Isaac, and T. Roy, "Affective Computing: Historical Foundations, Current Applications, and Future Trends," in *Emotions and Affect in Human Factors and Human-Computer Interaction*. Elsevier, 2017, pp. 213–231.

[19] R. Cowie, "The good our field can hope to do, the harm it should avoid," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 410–423, 2012.

[20] ——, *Ethical Issues in Affective Computing*. Oxford University Press, apr 2014.

[21] L. Devillers, "Human–Robot Interactions and Affective Computing: The Ethical Implications," *Robotics, AI, and Humanity*, pp. 205–211, 2021.

[22] S. Döring, P. Goldie, and S. McGuinness, "Principalism: A Method for the Ethics of Emotion-Oriented Machines," *Cognitive Technologies*, no. 9783642151835, pp. 713–724, 2011.

[23] A. F. Beavers and J. P. Slattery, "On the Moral Implications and Restrictions Surrounding Affective Computing," *Emotions and Affect in Human Factors and Human-Computer Interaction*, pp. 143–161, 2017.

[24] J. P. Sullins, "Robots, love, and sex: The ethics of building a love machine," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 398–409, 2012.

[25] M. Coeckelbergh, "Are emotional robots deceptive?" *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 388–393, 2012.

[26] M. Cooney, S. Pashami, A. Sant'anna, Y. Fan, and S. Nowaczyk, "Pitfalls of Affective Computing: How can the automatic visual communication of emotions lead to harm, and what can be done to mitigate such risks," in *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*. New York, New York, USA: Association for Computing Machinery, Inc, apr 2018, pp. 1563–1566.

[27] L. Stark and J. Hoey, "The ethics of emotion in artificial intelligence systems," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 782–793.

[28] T. Grote and O. Korn, "Risks and Potentials of Affective Computing. Why the ACM Code of Ethics Requires a Substantial Revision," in *In CHI 2017 workshop on Ethical Encounters in HCI*, 2017.

[29] R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, A. Nill Sánchez, D. Raji, J. Lisi Rankin, R. Richardson, J. Schultz, S. Myers West, and M. A. Whittaker, *AI Now 2019 Report*. AI Now Institute, 2019.

[30] A. McStay, V. Bakir, and L. Urquhart, *Emotion recognition: trends, social feeling, policy - Briefing paper: All Party Parliamentary Group on Artificial Intelligence*. The Emotional AI Lab, 2020.

[31] G. Greene, *The Ethics of AI and Emotional Intelligence: Data sources, applications, and questions for evaluating ethics risk*. Partnership on AI, 2020.

[32] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. M. Silenzio, and M. De Choudhury, "A taxonomy of ethical tensions in inferring mental health states from social media," in *Proceedings of the 2019 Conference on*

Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery, Inc, jan 2019, pp. 79–88.

[33] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, "Guidelines for human-AI interaction," in *Conference on Human Factors in Computing Systems - Proceedings*. New York, NY, USA: Association for Computing Machinery, may 2019, pp. 1–13.

[34] A. Batliner, S. Hantke, and B. W. Schuller, "Ethics and Good Practice in Computational Paralinguistics," *IEEE Transactions on Affective Computing*, pp. 1–1, sep 2020.

[35] A. McStay and P. Pavliscak, *Emotional Artificial Ingellience: Guidelines for Ethical Use*. The Emotional AI Lab, 2019.

[36] "Hollywood is tracking heart pounding movie scenes with wearable tech." https://www.wareable.com/wearable-tech/heart-racing-bear-scenes-the-revenant-2186

[37] "Millward Brown, Affectiva: A new way to test emotional responses to ads." https://www.bizcommunity.com/Article/224/19/70834.html

[38] "Emotional Movies - Netflix." https://www.netflix.com/browse/genre/4136

[39] "Mood and Genre filters for "Liked Songs" - The Spotify Community." https://community.spotify.com/t5/Community-Blog/Mood-and-Genre-filters-for-Liked-Songs/ba-p/5160106

[40] "AI Is Now Analyzing Candidates' Facial Expressions During Video Job Interviews - Inc.com." https://www.inc.com/minda-zetlin/ai-is-now-analyzing-candidates-facial-expressions-during-video-job-interviews.html

[41] "The robot-recruiter is coming — VCV's AI will read your face in a job interview - TechCrunch." https://techcrunch.com/2019/04/23/the-robot-recruiter-is-coming-vcvs-ai-will-read-your-face-in-a-job-interview/

[42] "7 Sentiment Analysis Tools to Improve Employee Engagement in 2020 - HR Technologist." https://www.hrtechnologist.com/articles/employee-engagement/sentiment-analytics-tools-features-price/

[43] "Driver Emotion Recognition and Real Time Facial Analysis for the Automotive Industry." https://blog.affectiva.com/driver-emotion-recognition-and-real-time-facial-analysis-for-the-automotive-industry

[44] "Affectiva and Nuance to Bring Emotional Intelligence to AI-Powered Automotive Assistants - Business Wire." https://www.businesswire.com/news/home/20180906005039/en/Affectiva-Nuance-Bring-Emotional-Intelligence-AI-Powered-Automotive

[45] "Kia & Hyundai introduce futuristic cars that can read your emotions." https://banyanhill.com/kia-hyundai-cars-that-read-emotions/

[46] "Tech Company Offers Law Enforcement and Gov't Agencies a Safe Lie Detector During COVID-19 Pandemic." https://converus.com/press-releases/tech-company-offers-law-enforcement-and-govt-agencies-a-safe-lie-detector-during-covid-19-pandemic/

[47] "The border guards you can't win over with a smile - BBC Future." https://www.bbc.com/future/article/20190416-the-ai-border-guards-you-cant-reason-with

[48] "DARPA Pays $1 Million For An AI App That Can Predict An Enemy's Emotions." https://www.forbes.com/sites/thomasbrewster/2020/07/15/the-pentagons-1-million-question-can-ai-predict-an-enemys-emotions/?sh=23c4cd7532b4

[49] "Aggression Detectors: The Unproven, Invasive Surveillance Technology Schools Are Using to Monitor Students." https://features.propublica.org/aggression-detector/the-unproven-invasive-surveillance-technology-schools-are-using-to-monitor-students/

[50] "Student Data Mining: Determining Demonstrated Interest." https://yourteenmag.com/teens-college/admissions/student-data-mining

[51] "Software that monitors students during tests perpetuates inequality and violates their privacy - MIT Technology Review." https://www.technologyreview.com/2020/08/07/1006132/software-algorithms-proctoring-online-tests-ai-ethics/

[52] "Rehearse your slide show with Presenter Coach - Office Support." https://support.microsoft.com/en-us/office/rehearse-your-slide-show-with-presenter-coach-cd7fc941-5c3b-498c-a225-83ef3f64f07b

[53] "Google Glass Is A Hit For Children With Autism." https://www.forbes.com/sites/johnnosta/2018/01/04/google-glass-is-a-hit-for-children-with-autism/?sh=ef0f31542042

[54] "Upgraded Google Glass Helps Autistic Kids "See" Emotions - IEEE Spectrum." https://spectrum.ieee.org/biomedical/bionics/upgraded-google-glass-helps-autistic-kids-see-emotions

[55] "Apollo Neuro - The Wearable Wellness Device For Stress Relief." https://apolloneuro.com/

[56] "Stress management - stress watch & monitoring - fitbit." https://www.fitbit.com/global/us/technology/stress

[57] "Sonde Health Detects Depression in the Sound of Your Voice." https://www.bostonmagazine.com/health/2016/07/14/sonde-health/

[58] "Facebook Increasingly Reliant on A.I. To Predict Suicide Risk : NPR." https://www.npr.org/2018/11/17/668408122/facebook-increasingly-reliant-on-a-i-to-predict-suicide-risk

[59] "Blueskeye AI - A spin-out from Nottingham Technology Ventures." https://www.nottinghamtechventures.com/companies/blueskeye-ai/

[60] "Robotic baby seals help COVID-19 patients with loneliness - The Burn-In." https://www.theburnin.com/technology/robotic-baby-seals-helping-covid-19-patients-loneliness-2020-6/

[61] "Nevermind, a Horror Game That Responds to Real-Life Fear and Stress - (The) Absolute." https://nevermindgame.com/

[62] "Cozmo - The Little Robot with a Big Personality - Anki Cozmo Robot." https://ankicozmorobot.com/

[63] "Call Centers Tap Voice-Analysis Software to Monitor Moods - WIRED." https://www.wired.com/story/this-call-may-be-monitored-for-tone-and-emotion/

[64] "Cogito raises $25 million to analyze phone calls with AI - VentureBeat." https://venturebeat.com/2020/11/18/cogito-raises-25-million-to-analyze-phone-calls-with-ai/

[65] "EyeSee store mannequins gather intelligence on shoppers." https://newatlas.com/eyesee-data-gathering-mannequins/25156/

[66] "Emotion recognition in retail: increasing sales in your store – Felenasoft." https://felenasoft.com/xeoma/en/articles/emotion_recognition_in_store

[67] "Emotional Art Gallery digital billboards cheer up anxious Stockholm commuters." https://www.dezeen.com/2019/03/25/emotional-art-gallery-stockholm/

[68] "AI wheelchair is controlled by a smile or other facial expressions." https://www.usatoday.com/story/tech/talkingtech/2018/12/03/ai-wheelchair-controlled-smile-other-facial-expressions/2183372002/

[69] J. A. Russell, "Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies," *Psychological Bulletin*, vol. 115, no. 1, p. 102, 1994.

[70] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychological science in the public interest*, vol. 20, no. 1, pp. 1–68, 2019.

[71] M. A. Arapova, "Cultural differences in Russian and Western smiling," *Russian Journal of Communication*, vol. 9, no. 1, pp. 34–52, jan 2017.

[72] R. Srinivasan and A. M. Martinez, "Cross-Cultural and Cultural-Specific Production and Perception of Facial Expressions of Emotion in the Wild," *IEEE Transactions on Affective Computing*, 2018.

[73] J. De Leersnyder, M. Boiger, and B. Mesquita, "Cultural regulation of emotion: individual, relational, and structural sources," *Frontiers in Psychology*, vol. 4, 2013.

[74] L. Rhue, "Racial Influence on Automated Perceptions of Emotions," *SSRN Electronic Journal*, dec 2018.

[75] J. A. Buolamwini, "Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers," Ph.D. dissertation, Massachusetts Institute of Technology, 2017.

[76] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 77–91.

[77] B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive Inequity in Object Detection," *arXiv e-prints*, p. arXiv:1902.11097, Feb 2019.

[78] N. Lim, "Cultural differences in emotion: differences in emotional arousal level between the East and the West," *Integrative Medicine Research*, vol. 5, no. 2, pp. 105–109, jun 2016.

[79] "Types of harm - Azure Application Architecture Guide - Microsoft Docs." https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/type-of-harm

[80] "Amazon ditched AI recruiting tool that favored men for technical jobs - The Guardian." https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

[81] "Universal declaration of human rights," United Nations, Dec. 1948.

[82] M. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation." *Optics express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.

[83] K. Boehner, R. Depaula, P. Dourish, and P. Sengers, "Affect : From Information to Interaction," in *Critical Computing: between Sense and Sensibility*, 2005, pp. 59–68.