The unbearable (technical) unreliability of automated facial emotion recognition

Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December: 1–17

DOI: 10.1177/20539517221129549
journals.sagepub.com/home/bds
SAGE

The unbearable (technical) unreliability of automated facial emotion recognition

Federico Cabitza[1,2], Andrea Campagner[1] and Martina Mattioli[1]

Abstract

Emotion recognition, and in particular acial emotion recognition (FER), is among the most controversial applications of machine learning, not least because of its ethical implications for human subjects. In this article, we address the controversial conjecture that machines can read emotions from our facial expressions by asking whether this task can be performed reliably. This means, rather than considering the potential harms or scientific soundness of facial emotion recognition systems, focusing on the reliability of the ground truths used to develop emotion recognition systems, assessing how well different human observers agree on the emotions they detect in subjects' faces. Additionally, we discuss the extent to which sharing context can help observers agree on the emotions they perceive on subjects' faces. Briefly, we demonstrate that when large and heterogeneous samples of observers are involved, the task of emotion detection from static images crumbles into inconsistency. We thus reveal that any endeavour to understand human behaviour from large sets of labelled patterns is over-ambitious, even if it were technically feasible. We conclude that we cannot speak of actual accuracy for facial emotion recognition systems for any practical purposes.

Keywords

Face emotion recognition, reliability, ground truth, annotation, survey, user study

Introduction

Emotional artificial intelligence (AI) (McStay, 2020) is an expression that encompasses all computational systems that leverage 'affective computing and AI techniques to sense, learn about and interact with human emotional life'. Within the emotional AI domain (but even more broadly, within the entire field of AI based on machine learning (ML) techniques), acial emotion recognition (FER),[1] which denotes applications that attempt to infer the emotions experienced by a person from their facial expression (Paiva-Silva et al., 2016; McStay, 2020; Barrett et al., 2019), is one of the most controversial (Ghotbi et al., 2021) and debated (Stark and Hoey, 2021) applications.

In fact, 'turning the human face into another object for measurement and categorization by automated processes controlled by powerful companies and governments touches the right to human dignity'[2] and 'the ability to extract […physiological and psychological characteristics such as ethnic origin, emotion and wellbeing…] from an image and the fact that a photograph can be taken from some distance without the knowledge of the data subject demonstrates the level of data protection issues which can arise from such technologies'.[3] On the other hand, opinions diverge among the specialist literature. Some authors highlight the accurate performance of FER applications and their potential benefits in a variety of fields; for instance, customer satisfaction (Bouzakraoui et al., 2019), car driver safety (Zepf et al., 2020), or the diagnosis of behavioural disorders (Paiva-Silva et al., 2016; Jiang et al., 2019). Others have raised concerns regarding the potentially harmful uses in sectors such as human resource (HR) selection (Mantello et al., 2021; Bucher, 2022), airport safety controls (Jay, 2017), and mass surveillance settings (Mozur, 2020). In addition, the scientific basis of FER applications has been called into question, either by equating their assumptions with pseudo-scientific theories, such as phrenology or physiognomy (Stark and Hutson, Forthcoming), or by questioning the validity of the

[1]University of Milan-Bicocca, Milan, Italy
[2]IRCCS Istituto Ortopedico Galeazzi, Milan, Italy

Corresponding author:
Federico Cabitza Dipartimento di Informatica, Viale Sarca 336, 20126 Milano, Italy
Email:federico.cabitza@unimib.it

reference psychological theories (Barrett et al., 2019), which assume the universality of emotion expressions through facial expressions (Elfenbein and Ambady, 2002). Lastly, others have noted that the use of proxy data (such as still and posed images) to infer emotions should be supported by other contextual information (McStay and Urquhart, 2019), especially if the output of the FER systems is used to make sensitive decisions, so as to avoid misinterpretation of the broader context. According to Stark and Hoey (2021) 'normative judgements can emerge from conceptual assumptions, themselves grounded in a particular interpretation of empirical data or the choice of what data is serving as a proxy for emotive expression'.

From a technical point of view, FER is a measurement procedure (Mari, 2003) in which the emotions conveyed in facial expressions are probabilistically gauged to detect the dominant one or a collection of prevalent emotions. As a result, FER can be related to the concepts of *validity* and *reliability*. A recognition system is *valid* if it recognizes what it is designed to recognize (i.e. basic emotions); it is *reliable* if the outcome of its recognition is consistent when applied to the same objects (i.e. a subject's expression). However, when FER is achieved by means of a classification system based on ML techniques, its reliability cannot (and should not) be separated from the reliability of its *ground truth*, i.e. *training* and *test* datasets (Cabitza et al., 2019). In this scenario, reliability is defined as the extent to which the categorical data from which the system is expected to develop its statistical model are generated from 'precise measurements', i.e. human 'recognitions' exhibiting an acceptable *agreement*. This is because, by definition, no classification model can outperform the quality of the human reference (Cabitza et al., 2020b).

In this study, we will *not* contribute to the vast (and heated) debate still currently going on about the *validity* of automatic FER systems (Franzoni et al., 2019; Feldman Barrett, 2021; Stark and Hoey, 2021), that is, we do not address the classification task from the conceptual point of view (how to define emotions, if possible at all) nor merely from the technical point of view (how to recognize emotions, whatever they are). For the sake of argument, we assume that the main psychological emotion models make perfect sense and we do not address how robust recognition algorithms are, how well they perform in external settings, and, most importantly, how useful they can be, i.e. whether they provide the benefits that their promoters envision and advocate.

Instead, we focus on the *reliability* of their ground truth, which is not a secondary concern from a pragmatic standpoint (Cabitza et al., 2020a, 2020b). To that end, we conducted a survey of the major FER datasets concentrating on their reported reliability as well as a small user study by which we address three related research questions: Do existing FER ground truths have an adequate level of reliability? Are human observers in agreement regarding the emotions they sense in static facial expressions? Do they agree more when the context information is shared before interpreting the expressions?

The first question is addressed in the 'Related work and motivations' section and the answer is in Table 3. The other questions are addressed by means of a user study described in the 'User study: Methods' section and whose results are reported in the 'Results' section. Finally, in the 'Discussion' section, we discuss these findings and their immediate implications, while in the 'Conclusion' section we interpret them within the bigger picture of FER reliability and relate them to implications for the use of automated FER systems in sensitive domains and critical human decision making.

Related work and motivations

In recent years there has been a rapid increase in interest in (and debate around) FER technologies, which have been used or proposed for use in a variety of settings (Kołakowska et al., 2014), including: in usability engineering, to detect usability issues (Johanssen et al., 2019); in behavioural therapy, to assist individuals with autism spectrum disorder in expressing and detecting emotions (Jiang et al., 2019; White et al., 2018); in computer-assisted car driving, to detect potentially dangerous emotional states (e.g. drowsiness, anger) (Jabbar et al., 2018; Zepf et al., 2020); in security and surveillance, to potentially prevent malicious behaviour (Bullington, 2005; Mukhopadhyay and Sharma, 2020); in HR, for assisting HR personnel in the interview and recruitment process (Vardarlier and Zafer, 2020; Bucher, 2022).

Despite the promising (but yet uncertified) results in the aforementioned application domains, or possibly because of them, the development of FER technology has been accompanied by criticism, on both ethical-legal and scientific-technological grounds (Franzoni et al., 2019; Crawford, 2021).

In terms of ethical and legal issues, numerous researchers and experts have highlighted the potential risks associated with the development and adoption of FER systems, including threats to privacy (Jay, 2017) and other fundamental human rights (Authors, 2020). Particularly shocking in this regard is the case of the Muslim minority Uyghurs in the Xinjiang region of China, who are allegedly subjected to daily surveillance with emotion detection cameras (Wakefield, 2021). Other related concerns include issues with image gathering and curation processes (Birhane and Prabhu, 2021), as well as copyright violations (Harvey and LaPlace, 2021). As a consequence, several authors have suggested the use of guidelines to avoid potential risks and infringements (Chancellor et al., 2019).

Table 1. Inter-rater reliability values and *p*-values for the first experiment.

Group	Multi-label	Distribution	Ordinal (Enjoyment)	Ordinal (Sadness)	Ordinal (Anger)	Ordinal (Disgust)	Ordinal (Contempt)	Ordinal (Surprise)	Ordinal (Fear)
No-context group	0.32	0.42	0.44	0.50	0.59	0.44	0.44	0.37	0.59
vs unacceptability	< 0.001*	< 0.001*	< 0.001*	< 0.001*	0.065	< 0.001*	< 0.001*	< 0.001*	0.032*
vs adequacy	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*
Context group	0.34	0.42	0.42	0.58	0.66	0.47	0.47	0.37	0.65
vs unacceptability	< 0.001*	< 0.001*	< 0.001*	0.013*	0.741	< 0.001*	< 0.001*	< 0.001*	0.570
vs adequacy	< 0.001*	< 0.001*	< 0.001*	< 0.001*	0.001*	< 0.001*	< 0.001*	< 0.001*	0.001*
Comparison *p*-value	0.033*	0.408	0.979	< 0.001*	0.001*	0.002*	0.019*	0.265	< 0.001*

* denote significance at the 95% confidence level.

Table 2. Intra-rater reliability values and *p*-values for the second experiment.

	Multi-label	Distribution	Ordinal (Enjoyment)	Ordinal (Sadness)	Ordinal (Anger)	Ordinal (Disgust)	Ordinal (Contempt)	Ordinal (Surprise)	Ordinal (Fear)
Value	0.41	0.41	0.62	0.69	0.76	0.60	0.57	0.56	0.69
vs unacceptability	< 0.001*	0.002*	0.639	0.348	0.093	0.222	0.088	0.140	0.649
vs adequacy	< 0.001*	< 0.001*	0.127	0.004*	0.336	0.013*	0.007*	0.017*	0.097

* denote significance at the 95% confidence level.

On the other hand, the technological viability of FER has been questioned in a variety of ways, ranging from mild stances that question the inability of the current approaches to take into account the subjectivity and context-dependence of emotion expression (Han et al., 2017; Stark and Hoey, 2021; Washington et al., 2021), to more hard-line stances that strongly reject the scientific soundness of the underlying psychological models (Barrett et al., 2019; Stark and Hutson, Forthcoming) or even refute the definition of emotions as measurable entities (Barrett, 2006).

As stated in the introduction, we are not going to discuss the viability of FER systems in relation to any of the above aspects. Instead, we focus on the *reliability* of their ground truth, i.e. the extent to which different human observers agree on the emotions they recognize in the face of some subjects. To better grasp these concepts and their significance, it should be underlined that current FER systems are based on ML and are thus trained on large labelled datasets of images. Human raters annotate these initially unlabelled datasets with one or more labels based on a specific emotion model: either categorical models, which describe emotions in terms of basic categories, e.g. Ekman's basic emotion model (Ekman, 1999); or dimensional models, which describe emotions in terms of continuous or ordinal feature vectors in a multi-dimensional space, e.g. the Valence-Arousal-Dominance (VAD) model (Mehrabian, 1996). The human raters tasked with annotating such datasets are typically given no information about the subjects involved, their social context, or the conditions under which such pictures were obtained (Barrett et al., 2019; Stark and Hoey, 2021). Because these datasets can contain tens of thousands of images, annotation is typically undertaken by multiple raters, each of whom may label multiple images (e.g., crowdsourcing settings). In this context, then *reliability* of the annotations refers to, and is operationalized as, the degree of agreement for the labels provided by the raters involved, i.e. what in the literature is usually called the *inter-rater reliability* (Hayes and Krippendorff, 2007).

Reliability and its measuring in FER

There are multiple measures for quantifying inter-rater reliability, including the simple percentage agreement P_o (whose use is however discouraged, as it does not take into account random agreement effects (Krippendorff, 2004)) and more statistically sound approaches, such as Cohen's and Fleiss' κ (Cohen, 1960; Fleiss, 1971), or Krippendorff's α (Krippendorff, 2018). We refer to the following sections, in particular, for the definition of Krippendorff's α, as it is the reliability measure we adopt in this article. In any case, any such measure of reliability thus represents and quantifies the intrinsic subjectivity and ambiguity of a task: the lower the value of a reliability measure, the higher the disagreement and degree of subjectivity of the task. Since ground truths used for training ML models are usually obtained by aggregating multiple labels (either into a single label, e.g. by majority voting or into a distribution), low-reliability values thus suggest the intrinsic ambiguous nature of the FER task, as well as the related risks of bias (Cabitza et al., 2020a).

Table 3. Summary and characteristics of the reviewed datasets.

	Number of images	Number of annotators	Reliability score	Number of subjects	Typology	Source	Emotion model
FER-2013 Goodfellow et al. (2013)	35,887 face images	1	N/A	N/A	Various people	Images from Google	categorical (7 expression labels)
EMOTIC Kosti et al. (2017)	23,571 images	1 (70% of images) 3 (20% of images) 5 (10% of images)	k = 0.30 (50% of images)	34320	Various people	Images from Google	categorical + dimensional (26 emotion labels + VAD)
Google FEC Vemulapalli and Agarwala (2019)	156K face images	6	N/A	N/A	Various people	Images from Flickr	categorical (triplets of emotion labels)
RAF-DB Li and Deng (2018)	30,000 face images	40	N/A	N/A	Various people	Images from the Internet	categorical (7 emotion labels)
AffectNet Mollahosseini et al. (2017)	450,000 face images	1 (92% of the images) 2 (8% of the images)	P_o = 0.607 (8% of images)	N/A	Various people	Images from the Internet	categorical (11 emotion labels)
LIRIS-CSE Khan et al. (2019)	26,000 frames	22	N/A	12	Ethnically diverse children	Movie clips of spontaneous expressions	categorical (6 emotion labels)
DEFSS Meuwissen et al. (2017)	404 face images	5 (at least)	N/A	116	Various people	Posed expressions	categorical (5 emotion labels)
DDCF Dalrymple et al. (2013)	80 images 5 angles 2 lighting conditions	20 (at least)	k = 0.780	80	Children	Posed expressions	categorical (8 emotion labels)
CAFE LoBue and Thrasher (2015)	1,192 images	100 raters per image on two occasions	N/A	154	Children	Posed expressions, lab setting	categorical (6 emotion labels)
EmoReact Nojavanasghari et al. (2016)	1,102 videos	3	$\alpha \in [-0.16, 0.64]$	63	Ethnically diverse children	Emotions elicited with YouTube videos	categorical + dimensional (8 emotion labels + VA)
NVIE Wang et al. (2010)	1,658 images	5	k = 0.65	215	Students	Posed and spontaneous expressions	categorical (7 emotion labels)
EMOTIONET Benitez-Quiroz et al. (2017)	1 million images	Automatically annotated with AUs	N/A	N/A	Various people	Images from the Internet	categorical (23 emotion labels)
SFEW Dhall et al. (2011)	700 frames	2	N/A	95	Various people	Frames extracted from movies	categorical (7 emotion labels)
TSINGHUA Yang et al. (2020)	1,128 images	60 (at least)	k = 0.761	110	Chinese people	Posed expressions	categorical (8 emotion labels)
EISVDB Wang et al. (2016)	810 speech videos	18	N/A	16	Chinese actors, elderly	Videos from TV series	categorical (7 emotion labels)
OMG-Emotion Barros et al. (2018)	7371 video clips	5	N/A	N/A	Various People	Videos from Youtube	dimensional (VA)
RADBOUD FACES DATABASE Langner et al. (2010)	4680 images	20	N/A	49 models	Dutch People	Photography taken in a studio	categorical (8 emotion labels)

(continued)

Table 3. Continued

	Number of images	Number of annotators	Reliability score	Number of subjects	Typology	Source	Emotion model
CASME II Yan et al. (2014)	247 images of micro-expressions	2	N/A	26	Various people	Emotions elicited with videoclips	categorical (5 emotion labels)

In the emotion model column, VAD stands for valence-arousal-dominance, while VA stands for valence-arousal.

An important question thus regards what levels of reliability are 'high enough' for a set of annotations to be considered sufficiently reliable to support further research. We describe two methods to address this question.

The first method is analytical and is based on the selection of a desired level of accuracy of the FER system. Once this value has been set, the nomogram depicted in Figure 1 can be used to establish a lower-bound threshold for adequate reliability, by following the relationship between reliability, ground truth correctness and ML model accuracy demonstrated in Cabitza et al. (2020a). Thus, having fixed a minimum acceptable value of actual accuracy *acc* (for a model whose measured level of accuracy is x), the selected reliability measure should be high enough to result in a ground truth quality g such that $x * g \geq acc$. For example, let us imagine that we need a FER system exhibiting $\geq$ 90% actual accuracy, and let us further assume that, by training the model on a ground truth assumed to be 100% correct (i.e. a universal assumption), we would be able to obtain a model whose measured accuracy (on a test set) in FER tasks is equal to 95%. The nomogram depicted in Figure 1 clearly shows that to achieve such a performance we would need a ground truth that is at least 96% correct. Thus, we set the minimum acceptable reliability score at $\alpha \sim 0.7$.

The second method builds on the body of knowledge that is available in the content analysis literature. Klaus Krippendorff was among the first researchers in the content analysis field to speculate how minimum acceptable coefficients should be chosen according to the importance of the conclusions to be drawn from annotated data and found the famous (and still wide spread) criteria proposed by Landis and Koch (1977) to be too broad and too overconfident. In 2004, Krippendorff (Krippendorff, 2004) suggested that when the costs of mistaken conclusions are high, as in biometric identification or other morally questionable AI applications such as FER, the minimum reliability value must also be set high. The recommendation of Krippendorff, followed and even reinforced by several other researchers (Carletta, 1996; Neuendorf, 2017), is that, lacking precise knowledge of the risks of drawing false conclusions from unreliable data, researchers should consider reliable data as those with reliability values higher than 0.8; should use data with values between 0.8 and 0.67 only to draw tentative conclusions; while data whose reliability measures are lower than 0.67 should be discarded.

Both methods provide similar reliability thresholds, in particular, the 0.7 threshold obtained by the analytic method is lower- and upper-bounded by the thresholds defined by Krippendorff. For this reason, in the following, we adopt these latter reliability reference values. It should be noted, however, that by the connection shown in 1, these two ground truth reliability thresholds can ultimately be translated into thresholds regarding the accuracy of a FER system, i.e. how many classification errors we are willing to accept in the face of a ground truth that has been built with a given level of agreement by the raters involved (assuming their representativeness and a uniform, natural, capacity for interpretation).

Nonetheless, despite its importance in quantifying the intrinsic complexity and subjectivity of any task, the reliability of the datasets commonly used to develop FER applications is often overlooked in the literature. There could be several explanations for this. For example, it has been widely noted that in ML research work on the model is seen as high-level and valuable, while *data work*, i.e. work on the underlying dataset, is typically devalued and considered to be mundane (Sambasivan and Veeraraghavan, 2022). Datasets can often be introduced in a few sentences, disregarding the source material, their creation, and provenance. Labour on the dataset (e.g. data cleaning) is often carried out by students, postdocs, or even crowdsourcing services, without any attention to the quality of the collected data (Paullada et al., 2021). In Table 3 we report the number of images, annotators per image, reliability scores (if reported), number of subjects (if reported), emotion classification model (either categorical models, which describe emotions in terms of basic categories, e.g., Ekman's basic emotion model, or dimensional models, which describe emotions in terms of continuous or ordinal feature vectors in a multi-dimensional space, e.g., the VAD model), typology and source for some of the most commonly used benchmark datasets for FER applications. Table 3 includes information on a selection of the datasets mentioned in two recent, comprehensive surveys (Dalvi et al., 2021; Mellouk and Handouzi, 2020), including the datasets which were openly available online and which are picture-based (i.e. no physiological data), and published after 2010.

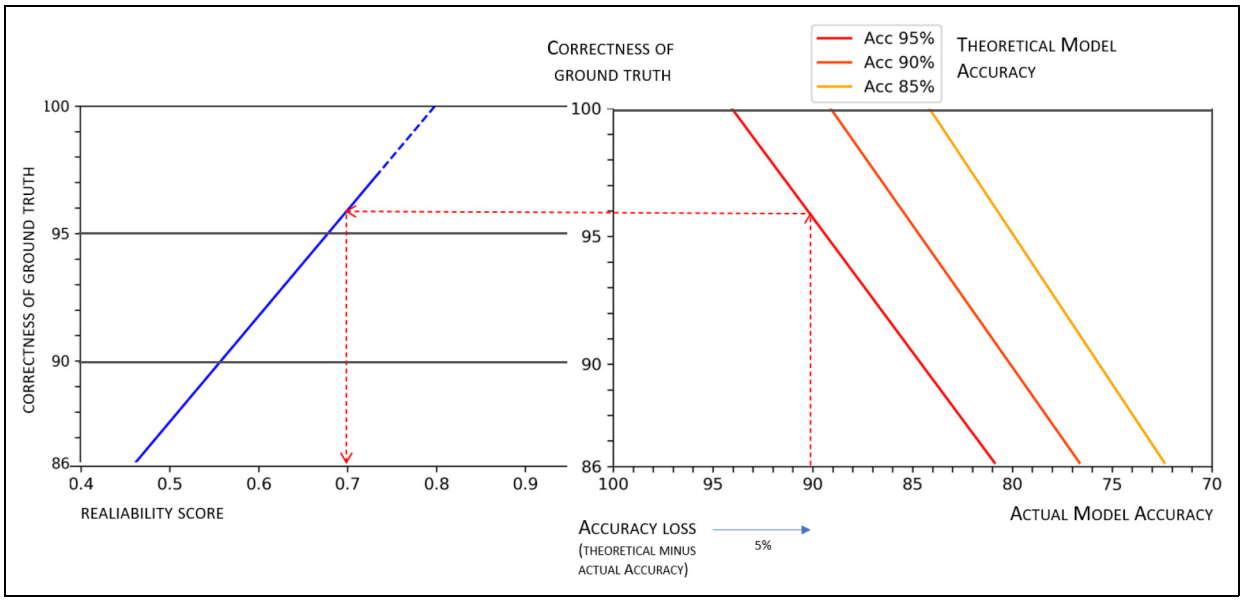

Figure 1. The relationship between reliability, correctness of the ground truth, and the actual accuracy of an machine learning (ML) model trained on that ground truth (adapted from Cabitza et al. (2020a)). The figure can be used as a sort of nomogram, i.e. a visual computation device that allows to approximate a function's computation by using only straight-line, equally graduated scales Given a minimum desirable level of accuracy (actual model accuracy) for an ML model and the corresponding theoretical model accuracy (i.e. the accuracy of the model as measured on a hypothetical 100% correct dataset), the diagram can be used to obtain the minimum acceptable reliability score for ground truth (cfr. the red dotted path). An example of use is shown in the figure with three dotted arrows and connected from right to left.

As highlighted in Table 3, only one third of the reviewed studies (i.e. 6 out of 16) reported reliability values. Furthermore, of the studies reporting the reliability of their ground truths, one study reported only the simple percentage agreement P_o, whose use to soundly assess the reliability (as mentioned above) has been discouraged (Krippendorff, 2004). Le Mau et al. (2021) similarly highlighted the use of the above-mentioned simple score as well as controversial (although widespread) reliability cutoff values (Barrett et al., 2019) (namely, no reliability for values below 0.2, weak reliability for values between 0.2 and 0.4, moderate reliability for values within the 0.4–0.7 range, high reliability for values above 0.7), with studies reporting mean reliability values of around 0.70. Even assuming that these datasets are not affected by defects and problems that have been detected in other datasets commonly used by the ML and deep learning community (Northcutt et al., 2021; Paullada et al., 2021), in all cases the reported values are lower than the adequate reliability (i.e. at least 0.8), and often (four datasets out of the six ones reporting reliability values) even lower than the above mentioned 0.67 threshold, which is lower than what is recommended as an acceptable threshold for supporting reliable research.

User study: Methods

Building on previous observations, the aim of this article was to evaluate the reliability of FER ground truths. Our analysis in Table 3 highlights the scarce reporting on the reliability of FER ground truths. It should also be highlighted that commonly used emotion recognition datasets do not usually provide the original multi-rater labellings, making replication or meta-validation studies hardly feasible. For our study, we thus considered a relatively small, but realistic, dataset of genuine (vs. posed) facial expressions, which is described in what follows. In particular, we designed and performed two annotation experiments to address the following three hypotheses:

1. *(H1)* Can the reliability of our FER ground truth be considered sufficient (according to the previously mentioned criteria) to support reliable research and analysis? In this sense, we will refer to the previously defined thresholds defined by Krippendorff (2004): one threshold below which labels should be discarded (*unacceptability threshold*), set at 0.67, and one threshold above which labels are of adequate reliability (*adequacy threshold*), set at 0.8.
2. *(H2)* Given that the raters involved in the annotation of FER ground truths are not usually provided with contextual information regarding the images to be annotated, does providing some sort of contextual information have any effect on the ground truth reliability? .
3. *(H3)* Even more problematic than different raters who differ in their interpretations of given facial expressions

is the same rater who cannot make up their mind about the same facial expressions after some (short) washout period, that is the lack of *intra-rater reliability*. We therefore also focus on whether the intra-rater reliability of our FER ground truth is high enough.

To address the hypotheses mentioned above, we involved a large number of raters (from now on, participants) in the annotation task of a FER dataset. The dataset encompassed 30 genuine, not posed, closeup pictures of facial expressions extracted by randomly selecting single frames from an online video depicting a conversation among participants in a class video meeting. The original video is freely available (upon registration) on Videvo,[4] and was released under a royalty-free, model released, free-use license. The pictures, each 300x300 pixel wide, depict five young subjects, four female students and one male student, at six different times, for a total of 30 pictures (see Figure 2).

First experiment: Inter-rater reliability

In the first experiment, we evaluated the inter-rater reliability of our FER ground truthing process. Participants in the first experiment were students enrolled in two master's degree courses at the University of Milano-Bicocca (namely, 'Interaction Design' and 'Digital Communication'), who had been invited to the experiment by direct e-mail after the rationale of the test had been explained in class. The experiment was conducted by means of an online questionnaire, implemented on the LimeSurvey platform,[5] through which the participants annotated the 30 pictures mentioned above. The participants were randomly assigned to two different groups:

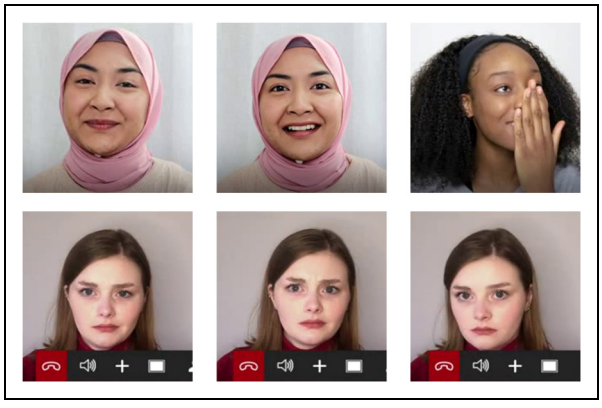

Figure 2. Six pictures depicting three subjects whose related emotions had to be recognized by the sample of raters involved in the study. The top images were associated with the highest agreement scores (easiest facial expressions to interpret and emotions to detect); the bottom ones with the lowest scores (hardest emotions to detect).

- Participants assigned to the first group (*no-context group*) were only shown the 30 pictures, one picture for each questionnaire page, randomly ordered.
- Participants assigned to the second group (*context group*), were first shown the high-definition (1138 × 640 pixels), 28-second original video from which the pictures had been extracted. The purpose was to provide the participants with some contextual information. The video was silent but clearly showing a professor teaching an online class, who eventually stops to wake up a student who has fallen asleep during his presentation, without reprimanding the student, but showing amused understanding. In fact, the teacher laughs it off whilst the other students giggle and joke, thus depicting a situation mainly characterized by levity, a mild sense of embarrassment, and fun. The video was played by the online platform on a loop, so that the participants could review it as many times as they wanted before filling out the questionnaire and annotating the subsequent images, just as the participants of the *no-context group* had to do with a sequence of pages as shown in Figure 3.

Participants were asked to annotate each image in the dataset according to Ekman's basic emotion model (Ekman and Friesen, 1986; Ekman, 1999), which is one of the categorical models most commonly adopted in FER applications (e.g., (Mehendale, 2020; Brodny et al., 2016; Stark, 2018)) . According to this model, each facial expression can be associated with one or more basic emotions from among 7 basic types (Ekman and Friesen, 1986) conjectured to be universal across human cultures: namely, *enjoyment, contempt, surprise, fear, sadness, disgust* and *anger* .[6] In order to also take into account dimensional models, in our experimental setting raters were asked to indicate, for each image, a sort of arousal,[7] expressed in terms of the perceived pertinence of each basic emotion for the facial expression therein depicted, by means of a 5-value ordinal scale ranging from 1 (this emotion cannot be detected in the current expression) to 5 (this emotion is present and it's very intensely expressed).[8] This scale thus produces both categorical annotations and dimensional (continuous) ones, as shown in the following paragraphs.

After collecting the annotations, we measured the reliability of the obtained multi-rater labels through Krippendorff's α metric Krippendorff (2004). The intuitive definition of reliability is simple: to what extent can we rely upon an agent's decisions? Similarly, to what extent can we rely on data to train a predictive model and for it to produce actual predictions? Despite the broad scope of this concept Cabitza et al. (2020b), we focus on the metrological interpretation of reliability, which is concerned with measurement precision and, more broadly, consistency of rating and labelling: for example, raters who apply the same label to the same

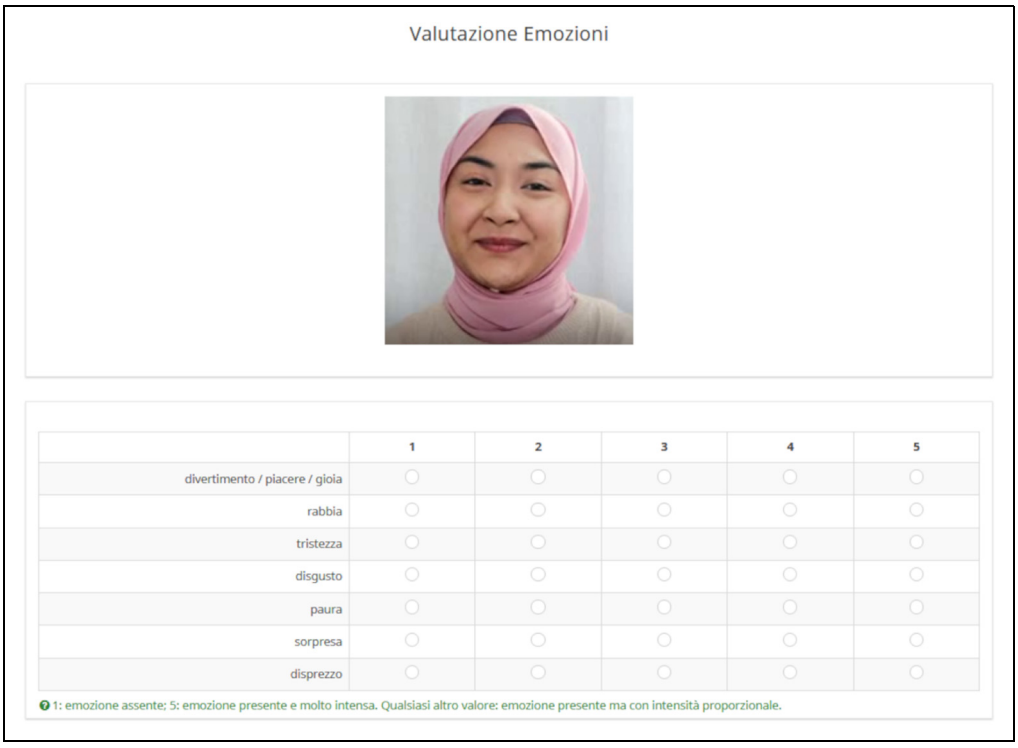

Figure 3. Screenshot of an annotation page of the online questionnaire used during the experiment. The legend on the bottom runs (in Italian): '1: emotion absent; 5: emotion present and very intense. Any other value: emotion present but with proportional intensity.'

case. In fact, we focus on the more technical notion of reliability as the complement of observer variability Krippendorff (2004), either between multiple annotators (inter-rater) or for single raters (intra-rater consistency). In this sense, assessing reliability means evaluating the degree to which the observed agreement among raters is genuine and not due to chance. More in detail, the formal definition of Krippendorff's α is:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

where $D_o = \frac{1}{30}\sum_{j=1}^{30} \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{i'=i+1}^{k} \delta(r_i^j, r_{i'}^j)$ is the observed agreement, $\delta(r_i^j, r_{i'}^j) \in [0, 1]$ is a distance function representing the closeness of two different ratings and D_e is the expected agreement score under the empirical distribution. To better illustrate the notion of inter-rater reliability, here we report the pictures associated with the least and highest reliability in Figures 4 and 5 respectively.

We considered, in particular, three different representations of ratings, which include some of the most common label representation approaches in the FER literature (Ko, 2018; Washington et al., 2021):

- Multi-label representation: each rater judgement is expressed in terms of the emotion(s) that they considered more intense, discarding the other less intense ones. For instance, for rater i, the list of emotions 'surprise' and 'enjoyment'.
- Distribution-based representation: each rater judgement is represented in percentage terms of a whole constituted by all the non-absent emotions, like, for rater i the list: 'enjoyment': 67%', 'surprise': 23%', 'anger': 10%'.
- Ordinal representation: each rater judgement is simply expressed as the list of reported emotional intensity values between 1 and 5.

Clearly, the multi-label representation corresponds to a categorical emotion model, while the ordinal and distribution-based representations take into account features of both categorical and dimensional emotion models. These three different representations are reflected in a different definition of the δ distance function in Krippendorff's α. More specifically:

- For the multi-label representation, the distance between two ratings is defined as the intersection-over-union distance, that is two ratings are considered more similar if they encompass the same emotions.
- For the distribution-based representation, the distance between two ratings is defined as the Euclidean distance between the respective probability distributions.

Figure 4. One of the pictures associated with the lowest reliability scores. For all basic emotions excluding 'enjoyment' the ratings spread the range of emotional intensity.

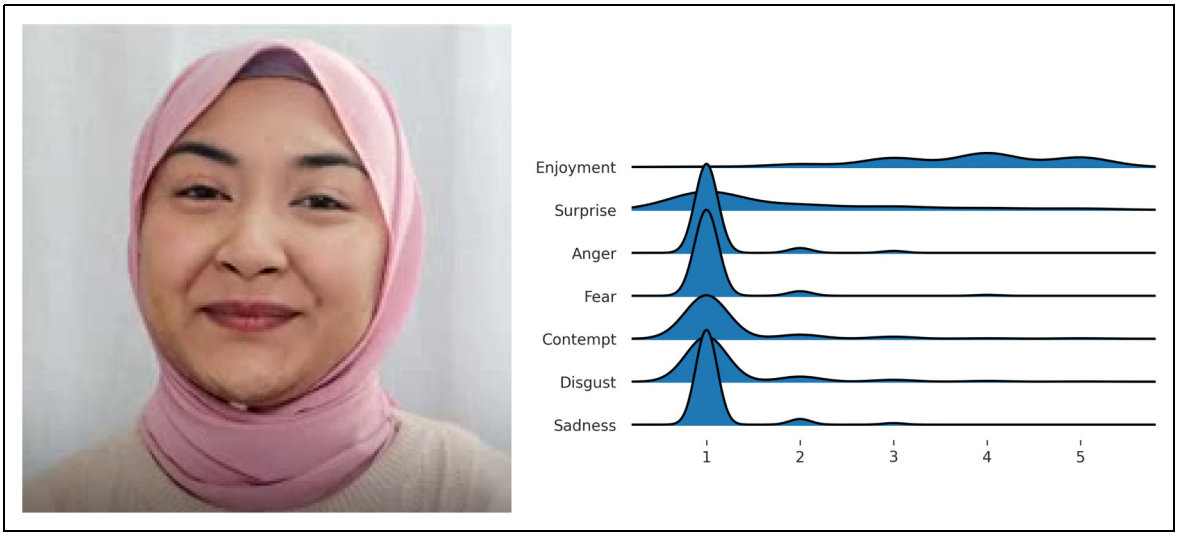

Figure 5. One of the pictures associated with the highest reliability scores. Note the peak on emotion intensity 1 (i.e.'absent') for all basic emotions except 'enjoyment'.

- For the ordinal representation, the distance between two ratings is simply defined, for each emotion e, by the normalized difference between the two ratings.

We, therefore, considered nine different values of reliability (one for each basic emotion, in addition to the multi-label and distribution-based ones). For each of these values we evaluated the two hypotheses ($H1$) and ($H2$), that is:

- ($H1$): Is the reliability of either of the two groups sufficiently high? For this, we adopted the two previously defined thresholds, and then compared the reliability values observed in both the above groups against these thresholds, by means of the one-sample Student's t test.
- ($H2$): Is there a significant difference between the reliability for participants in the *no-context group* and participants in the *context group*? In particular, is the reliability for the *context group* significantly greater? To test ($H2$), we compared the reliability values for the two groups, by means of Wilcoxon signed-rank test.

In both cases, we considered p-values lower than 0.05 to be significant (at a 95% confidence level).

Second experiment: Intra-rater reliability

In the second experiment, we evaluated the *intra-rater reliability* in emotion recognition ground truthing, in order to answer our third hypothesis ($H3$).

Participants in the second experiment were not the same students as those involved in the first one, however, they were from the same classes, and they were tasked with filling in a slightly modified version of the questionnaire shown to the *context group*. As in the context group cohort in the first experiment, all participants were first shown the video from which the pictures were extracted. The rationale for this was that we aimed to evaluate intra-rater reliability in the most conservative scenario, in which the raters had access to all relevant contextual information for the labelling task. However, unlike the first experiment,from the 30 pictures shown in the questionnaire, 5 were repeated (not consequently, and at a random number of pictures apart, to minimize the likelihood that participants might notice the repetition), so that only 25 expressions were shown and the pairs of pictures could be used to evaluate the intra-rater reliability.

As in the first experiment, intra-rater reliability was evaluated by means of Krippendorff's α, with the same nine representation models as previously described. For each pair of repeated images and each rater, we evaluated the agreement between the ratings given by the same rater for the two images. The intra-rater reliability was then computed by averaging the pair-wise reliability values. To better illustrate the notion of intra-rater reliability, here we report one of the pictures exhibiting the lowest intra-rater reliability score, in Figure 6.

After computing the reliability values, we tested whether the intra-rater reliability was sufficiently high, by comparing the 9 reliability α values against the *unacceptability* and *adequacy* thresholds, by means of the one-sample Student's t test. P-values lower than 0.05 were considered significant.

Results

After closing the survey for the first experiment, we collected a total of 198 complete responses, from as many participants, for a total of 5940 expressions and 41,580 emotion ratings. The difference in rating distributions for the two groups is reported in Figure 7.

A total of 101 participants were assigned to the *no-context group*, while 97 participants were assigned to the *context group*. The reliability values, for each of the representation models and the two groups, are reported in Table 1.

All reliability values were significantly lower than the *adequacy threshold*. In addition, all values except for the ordinal representation for 'anger' (for both groups) and 'fear' (for the *context group*) were significantly lower than the *unacceptability threshold*. The reliability values for the *context group* were higher than the corresponding values for the *no-context group* for all representation models except for the distribution-based model and the ordinal model for the enjoyment and surprise emotions. The distribution of emotion ratings (for 'enjoyment' and 'disgust') for the picture for which the difference in reliability scores between the two groups was highest is reported in Figure 8.

With regard to the second experiment, we collected a total of 51 complete responses, from as many participants for a total of 1530 expressions, and 10,710 emotion ratings. The reliability values are reported in Table 2.

The reliability values for the multi-label and distribution-based representations were significantly lower than the *unacceptability threshold*, while all other reliability values

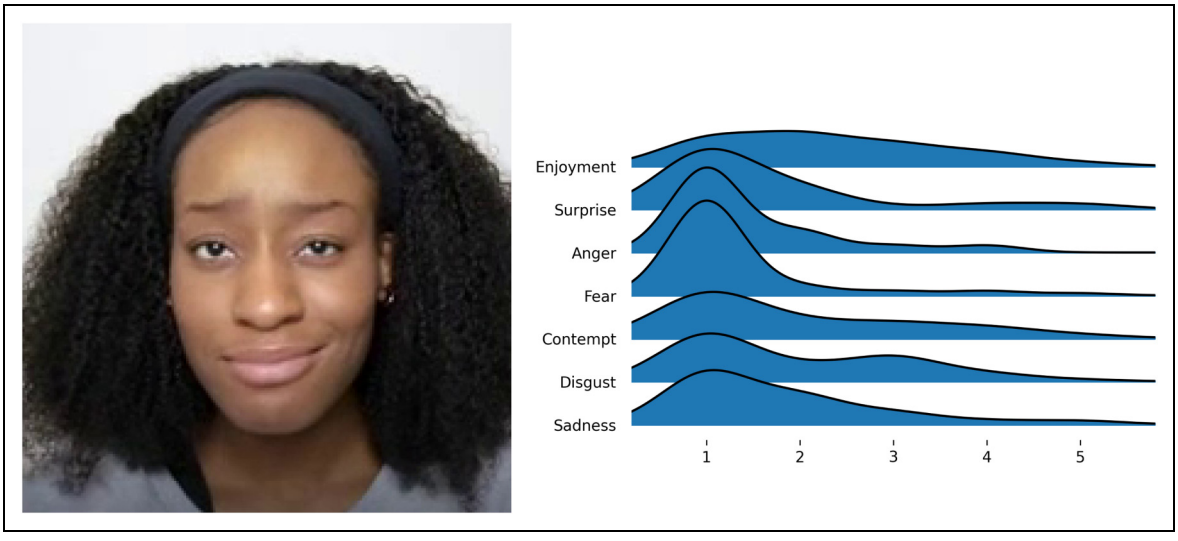

Figure 6. One of the pictures with the lowest intra-rater reliability scores. Note the large spread over the whole emotion intensity range, with a peak around intensity 1 (i.e. absent) for all basic emotions.

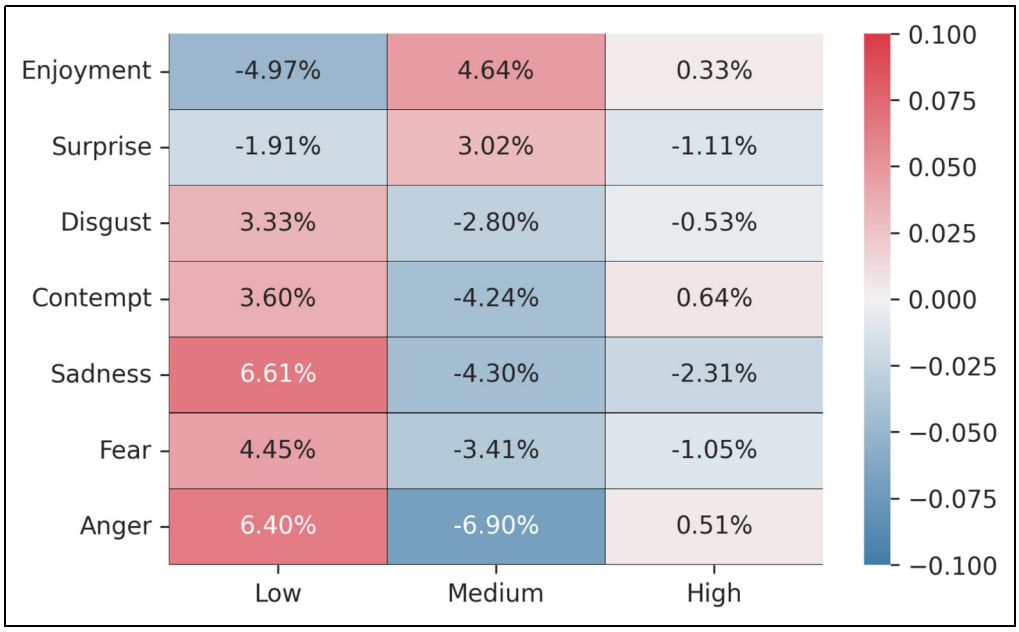

Figure 7. Differences in emotion ratings between the *context* and *no-context* groups. The reported values correspond to the frequency for *context group*, minus the frequency for the *no-context group*. Red cells, therefore, denote a higher frequency for the *context group* for the respective rating and emotion, while blue cells denote a higher frequency for the *no-context group*. Note the higher frequencies in the medium and high-intensity values for 'enjoyment', as well as for the low intensity (i.e. rating 1) for other emotions (except 'surprise'), for the *context group*.

were not significantly different from the former threshold. In particular, the reliability values for the emotions of 'sadness', 'anger' and 'fear' were higher than the *unacceptability threshold*. By contrast, all reliability scores were lower than the *adequacy threshold*, and this difference was significant for all representation models except the ordinal one for the emotions of 'enjoyment', 'anger' and 'fear'.

Discussion

As Table 3 shows, the FER community seems to suffer from a problem of ground truth reliability, i.e. a problem with the reliability of the data used to train their classification systems. In fact, all of the reliability scores are below the adequacy threshold, and two-thirds of the reported ones are even below the unacceptability threshold defined in the previous section. However, low ground truth reliability is a much smaller problem than ignoring it (as the 'Reliability Score' column in Table 3 is empty for 10 datasets out of 16), as this implies not taking countermeasures or, worse still, taking majority subjective opinions on facial expressions as real objective phenomena, to ground the delivery of multiple delicate services (such as profiling, diagnosis) on.

There may be several reasons for the low awareness and underestimation of this reliability problem, including the devaluation of data work in the ML community, or the methodological decision to adopt outdated criteria that overstate rater agreement and underestimate observer variability (Landis and Koch, 1977). In this study, we assumed that data collected by people who agreed about what expression or emotion they were observing less than two out of three times (or 7 out of 10), not including the number of times they did so by chance, cannot be trusted, especially in sensitive contexts or for applications with significant legal effects. The ongoing popularity of FER systems and similarly controversial applications of ML in the real-world often side-steps relevant discussions (Bender, 2022) about the validity, and even the semantics, of the underlying data because of the promise of knowing what customers 'really feel' (Munn, 2020). In stark contrast with this, we believe that these untrustworthy data cannot be used to develop systems that will end up being arbitrary, yet disguised as an objective evaluation (Basile et al., 2021). In our view, systems based on these data are not capable of producing useful identifications and recognitions of what emotions people actually feel. Our study, although limited with regard to the number of depicted subjects (6) and pictures (30) considered, thus contributes to the literature that converges to this conclusion. It also provides commercial stakeholders with further empirical evidence backing their decision to discontinue facial analysis screening, as recently done by HireVue, Inc. (Bucher, 2022).

In terms of our first hypothesis ($H1$) (i.e. whether the inter-rater agreement about perceived emotions expressed

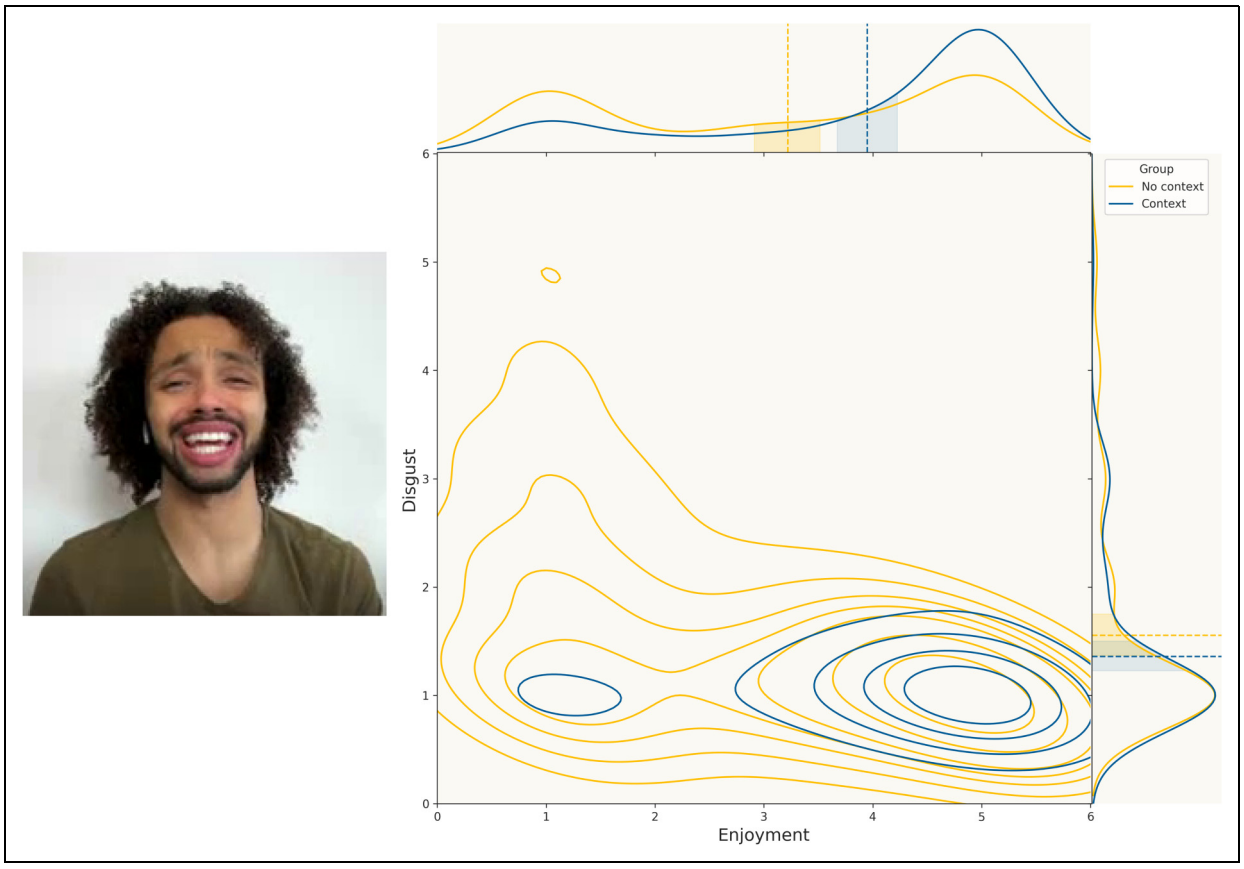

Figure 8. Distribution of emotion ratings (for 'enjoyment' and 'disgust') for the image for which the difference in reliability scores between the two groups was highest.

in facial expressions is sufficient for high-risk applications in sensitive domains), we have shown that human observers do not agree sufficiently often about the emotions that they perceive on the basis of still images. In fact, as highlighted by the results reported in Table 1, in all cases the reliability values were significantly lower than the *adequacy threshold*, and in only 3 cases out of 18 (16%) were the observed values higher than the *unacceptability threshold*. It should also be noted that this observation holds for both the *context* and *no-context group*s, as further highlighted in Figures 4 and 5, which depict two pictures with the lowest and highest reliability scores respectively. While the participants from the two groups strongly agreed about which emotions should be excluded (i.e. 'enjoyment' in Figure 4 and all other emotions in Figure 5), the ratings for the other emotions were more uniformly distributed, denoting a lack of agreement among the participants.

Our second hypothesis, (*H*2) (i.e. whether providing the raters with additional context would increase the observed reliability), on the other hand, was motivated by the idea that at least part of the reliability problem, mentioned above and which we also observed, may be related to a lack of context. In fact, as also noted by Ekman himself, 'emotions are a process' (Goldie, 2000), rather than discrete units or emotional states, and even what we recognize as 'faces', rather than physical entities, have been recently reconceptualized as 'distributed accomplishments' (Bucher, 2022). It is also well known that the interpretation of emotions from still images is affected by the so-called Kuleshov effect (Barratt et al., 2016), that is the feeling that a shot conveys to the viewer is significantly influenced by the preceding and following shots. That notwithstanding, automatic FER is usually based on still images only (Paiva-Silva et al., 2016), most of the time posed (and hence potentially overloaded) portraits, or alternatively frames extracted from staged videos, which are therefore de-contextualized and associated with a limited set of categories (i.e. the basic emotions). As also recently observed (Paiva-Silva et al., 2016), the validity of static human face stimuli is the object of criticism in the scholarly community studying facial expression, thus motivating a growing interest toward so-called *appraisal*-based emotional AI (McStay and Urquhart, 2019), which also takes into account contextual and physiological information.

In our study, we collected evidence that providing context, in the form of a short silent video depicting the

situation where the facial expressions were produced, can yield significant differences, making the agreement among observers higher, and hence the reliability sounder, although still insufficient for reliable FER systems. Indeed, in 6 cases out of 9 (67%) the reliability observed in the *context group* was significantly higher than for the *no-context group*. Furthermore, even though the reliability values were small for both groups, two out of the three reliability values which were greater than the *unacceptability threshold* occurred for the *context group*.

Further details in regard to the results can be observed from Table 1, which reports the differences in the frequencies of emotions' intensity values for the two groups. The participants in the *context group* were better able to identify the dominant emotions (that is, 'enjoyment' and 'surprise') in the original video, as this can be noticed by the negative rates in the low-intensity values as well as the positive rates in the medium (and high, for 'enjoyment') intensity values. Similarly, the participants in the *context group* were better able to exclude the less-pertinent emotions, as can be noticed by the positive rates in the low-intensity values for all other emotions. Thus, access to context information not only improved the reliability of the raters, but it also improved their ability to identify the emotions expressed. The same point is highlighted in Figure 8, which represents one of the pictures for which the difference in reliability between the two groups was larger. Indeed, we can easily note how the distribution of ratings from participants in the *context group* was much more concentrated on the average and high emotional intensity values for the emotion 'enjoyment', as well as on the low-intensity values for the emotion 'disgust'. By contrast, the ratings expressed by the *no-context group* were much more uniformly distributed across the whole range of emotional intensity.

Finally, the answer to our third hypothesis ((*H*3), i.e. whether the intra-rater reliability in our FER ground truth was sufficiently high) is negative, albeit less significantly so. In particular, while most (7 out of 9, 78%) intra-rater reliability values were higher than the *unacceptability threshold*, 6 out of 9 (67%) values were significantly lower than the *adequacy threshold*. As shown in Figure 6, pictures with low intra-rater reliability were characterized by a distribution that was relatively uniformly spread throughout the low emotional intensity range. In other words, the emotion ratings for these pictures highlight the uncertainty of the raters, as they were not able to select the most representative emotion but rather reported low intensity for all basic emotions.

Conclusion

Summarizing our findings, we note that our results are consistent with previously reported low-reliability values for FER ground truths (see Table 3). Although this fact was observed in previous studies (Stark and Hoey, 2021), we corroborated this claim by performing a reliability-focused literature survey and, mostly important, by making the point that the reported low-reliability score is excessively low for a viable use of that information. This means that excessively low reliability on FER data necessarily entails low accuracy of FER applications (Cabitza et al., 2020a), as shown in Figure 1 and discussed in the 'Related work of motivations' section.

We also confirmed that access to situational context improves emotion recognition by humans (in terms of reduced disagreement), and this kind of context is what it is usually missing in the reference data that are used to train automatic recognition systems (besides often lacking also the naturalness that cannot be represented in posed stills), not to speak of how AI systems could actually understand situational context even once this was supplied to them.

Our findings thus provide empirical support for the inherent subjectivity and ambiguity of FER tasks, which have been already discussed in psychological, neurological and anthropological studies (Abu-Lughod and Lutz, 1990; Barrett et al., 2019; LeDoux and Hofmann, 2018; Wearne et al., 2019), but seldom related to its potential impact on the development of FER technologies. As a consequence, our results challenge the technical reliability and soundness, as well as the very definition of accuracy (Cabitza et al., 2021), of FER systems based on ground truths obtained by aggregating annotations provided by multiple raters. Indeed, low-reliability values correspond to a low agreement between the annotating raters: therefore, any FER system trained on such aggregated ground truths (even highly accurate ones) would be able to identify only a part of the emotions associated with the facial expressions to be classified.

Lastly, our results regarding intra-rater reliability also question the consistency and reliability of the individual annotations: although the intra-rater reliability was higher than the inter-rater reliability (indeed, most values were higher than the *unacceptability threshold*), it still failed to meet the requirements of good reliability (i.e. being significantly greater than the *adequacy threshold*).

We believe these observations lend further support to recent calls for adopting alternative annotation practices, and related ML methods, such as *perspectivist ground truthing* (Basile et al., 2021), which take into account all available annotations (to avoid the problem of low inter-rater reliability), as well as additional information about the raters, such as their confidence or uncertainty (to address the problem of low intra-rater reliability). More in general, our results resonate with recent initiatives (Bender and Friedman, 2018; Gebru et al., 2021; Holland et al., 2018) aimed at raising awareness about the data production process (Gitelman, 2013), including the need to document in which (technical, social, economical and

political) context the data were collected and how annotation was actually performed. As discussed in a recent survey (Paullada et al., 2021) and as we highlighted in the previous sections, the 'data' aspect has always been a critical aspect of ML development but it remains extensively mishandled in practice and ignored in theory. We believe that a shift in focus from model development to the issues and approaches mentioned above could allow researchers to develop FER systems that are more representative of the subjectivity of the task.

All that said, this is why we assert the somehow provocative (but grounded) claim that *we cannot speak of accuracy for facial expression and emotion recognition technology*: in fact, no reference can be reliably established against which to compute meaningful error rates. One could object that the present study regards only one particular FER dataset, which nevertheless was built by involving a large number of raters, much larger than in common facial expression datasets (see Table 3), and one particular set of emotion labels (the Ekman's basic emotion model), but we feel that this interpretation of our study would be too narrow. Indeed, in Table 3 we show several reference datasets as examples of reliability scores which when reported, are *low*, and, through the nexus demonstrated in Cabitza et al. (2020a) and depicted in Figure 1, that low reliability entails low accuracy.

This means that, besides any ethical considerations (Stark and Hoey, 2021; Ghotbi et al., 2021), the irredeemable low reliability of emotion classification poses important challenges to the validation, and hence certification, of FER technologies or of the systems embedding FER capabilities. We believe these difficulties are especially relevant due to the growing interest in so-called *appraisal-based* FER systems McStay and Urquhart (2019), i.e. systems that rely not only on still images and basic emotion categories but also on multi-faceted contextual, physiological or personal information. While we showed that using additional contextual information (such as videos) could improve the reliability of the data underlying such applications, obtaining such information clearly poses even greater ethical and privacy risks for the individuals involved. Furthermore, in our experiments, we showed how even the emotion ratings produced with the aid of more informative contextual information were associated with reliability that could be deemed insufficient to enable practical applications.

For this reason, we share the appeal recently made by Ienca and Malgieri (2021) that the next regulations on Artificial Intelligence, and among these the European Artificial Intelligence Act, should explicitly include the AI systems that rely on mental information derived from emotion recognition systems in the high-risk list. Regulations should subject these systems to specific compliance duties and requirements to manage the risks involved, such as conformity certifications, risk management plans, and human oversight.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Federico Cabitza https://orcid.org/0000-0002-4065-3415
Andrea Campagner https://orcid.org/0000-0002-0027-5157

Notes

1. In the specialized literature this acronym is currently used to designate two similar expressions: FER and facial expression recognition; we adopt the former one because we understand that the recognition of facial expressions by digital systems is always related to, or instrumental to, the recognition of emotions expressed by facial expression and the use of this latter information to adapt the behaviour/response of the digital systems (on this see also (Ko, 2018)).
2. as stated by Europe's data protection authority and cited in (McStay, 2020)
3. Article 29 Data Protection Working Party (2012) Opinion 3/2012 on developments in biometric technologies. Available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2012/wp193˙en.pdf (accessed 8 April 2022).
4. https://www.videvo.net/video/professor-wakes-up-student-during-online-class/678555/. Last accessed on the 20th of Janaury, 2022. Static screenshot archived at https://archive.is/wip/4TQtD.
5. https://www.limesurvey.org/
6. Also an eighth emotion can be considered, i.e. *neutral*, characterized by the absence of any of the previous ones
7. Indeed, arousal is the shared dimension in the most common dimensional models, namely valence-arousal-dominance (Nicolaou et al., 2011) and pleasure-arousal-dominance (Mehrabian, 1996)
8. Technically, this scale is a direct derivation of the Positive and Negative Affect Schedule scale, which is very popular in affect measuring Adams et al. (2013).

References

Abu-Lughod L and Lutz CA (1990) Introduction: Emotion, discourse, and the politics of everyday life. *Language and the Politics of Emotion* 1: 1–23.

Adams S, Penton-Voak IS, Harmer CJ, et al. (2013) Effects of emotion recognition training on mood among individuals with high levels of depressive symptoms: Study protocol for a randomised controlled trial. *Trials* 14(1): 1–7.

Authors V (2020) Emotional entanglement: China's emotion recognition market and its implications for human rights. Technical report, ARTICLE 19.

Barratt D, Rédei AC, Van de Weijer J, et al. (2016) Does the kuleshov effect really exist? revisiting a classic film experiment on facial expressions and emotional contexts. *Perception* 45(8): 847–874.

Barrett LF (2006) Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review* 10(1): 20–46. doi:10.1207/s15327957pspr1001_2. PMID: 16430327

Barrett LF, Adolphs R, Marsella S, et al. (2019) Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest* 20(1): 1–68.

Barros P, Churamani N, Lakomkin E, et al. (2018) The omg-emotion behavior dataset. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–7.

Basile V, Cabitza F, Campagner A, et al. (2021) Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.

Bender EM (2022) Look behind the curtain: Don't be dazzled by claims of 'artificial intelligence'. *The Seattle Times*.

Bender EM and Friedman B (2018) Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6: 587–604.

Benitez-Quiroz CF, Srinivasan R, Feng Q, et al. (2017) Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210*.

Birhane A and Prabhu VU (2021) Large image datasets: A pyrrhic win for computer vision? In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 1536–1546.

Bouzakraoui MS, Sadiq A and Alaoui AY (2019) Appreciation of customer satisfaction through analysis facial expressions and emotions recognition. In: *2019 4th World Conference on Complex Systems (WCCS)*. IEEE, pp. 1–5.

Brodny G, Kolakowska A, Landowska A, et al. (2016) Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions. *2016 9th International Conference on Human System Interactions (HSI)* : 397–404.

Bucher T (2022) Facing ai: conceptualizing 'faice communication'as the modus operandi of facial recognition systems. *Media, Culture & Society*: 01634437211036975.

Bullington J (2005) 'affective' computing and emotion recognition systems: the future of biometric surveillance? In: *Proceedings of the 2nd annual conference on Information security curriculum development*. pp. 95–99.

Cabitza F, Ciucci D and Rasoini R (2019) A giant with feet of clay: On the validity of the data that feed machine learning in medicine. In: *Lecture Notes in Information Systems and Organisation, vol. 28*, volume 28. pp. 121–136. doi:10.1007/978-3-319-90503-7_10.

Cabitza F, Campagner A, Albano D, et al. et al (2020a) The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability. *Applied Sciences* 10(11): 4014.

Cabitza F, Campagner A and Sconfienza LM (2020b) As if sand were stone. new concepts and metrics to probe the ground on which to build trustable ai. *BMC Medical Informatics and Decision Making* 20(1): 1–21.

Cabitza F, Campagner A and Datteri E (2021) To err is (only) human. reflections on how to move from accuracy to trust for medical ai. In: *Exploring Innovation in a Digital World*. Springer, pp. 36–49.

Carletta J (1996) Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22(2): 249–254.

Chancellor S, Birnbaum ML, Caine ED, et al. (2019) A taxonomy of ethical tensions in inferring mental health states from social media. In: *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255, p. 79–88. doi:10.1145/3287560.3287587.

Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1): 37–46.

Crawford K (2021) Artificial intelligence is misreading human emotion. *The Atlantic* https://www.theatlantic.com/technology/archive/2021/04/artifi.

Dalrymple KA, Gomez J and Duchaine B (2013) The dartmouth database of children's faces: Acquisition and validation of a new face stimulus set. *PloS one* 8(11): e79131.

Dalvi C, Rathod M, Patil S, et al. (2021) A survey of ai-based facial emotion recognition: Features, ml & dl techniques, age-wise datasets and future directions. *IEEE Access* 9: 165806.

Dhall A, Goecke R, Lucey S, et al. (2011) Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, pp. 2106–2112.

Ekman P (1999) Basic emotions. *Handbook of Cognition and Emotion* 98(45-60): 16.

Ekman P and Friesen WV (1986) A new pan-cultural facial expression of emotion. *Motivation and Emotion* 10(2): 159–168.

Elfenbein HA and Ambady N (2002) On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin* 128(2): 203.

Feldman Barrett L (2021) Debate about universal facial expressions goes big. *Nature*.

Fleiss JL (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5): 378.

Franzoni V, Vallverdù J and Milani A (2019) Errors, biases and overconfidence in artificial emotional modeling. In: *IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume*. pp. 86–90.

Gebru T, Morgenstern J, Vecchione B, et al. (2021) Datasheets for datasets. *Communications of the ACM* 64(12): 86–92.

Ghotbi N, Ho MT and Mantello P (2021) Attitude of college students towards ethical issues of artificial intelligence in an international university in japan. *AI & SOCIETY*: 1–8.

Gitelman L (2013) *Raw Data is An Oxymoron*. Cambridge, MA: MIT press.

Goldie P et al (2000) *The Emotions: A Philosophical Exploration*. Oxford, UK: Oxford University Press.

Goodfellow IJ, Erhan D, Carrier PL, et al. et al (2013) Challenges in representation learning: A report on three machine learning contests. In: *International conference on neural information processing*. Springer, pp. 117–124.

Han J, Zhang Z, Schmitt M, et al. (2017) From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In: *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17. New York, NY, USA: Association for Computing Machinery. ISBN 9781450349062, p. 890–897. doi:10.1145/3123266.3123383. https://doi.org/10.1145/3123266.3123383.

Harvey A and LaPlace J (2021) Exposing.ai. https://exposing.ai.

Hayes AF and Krippendorff K (2007) Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures* 1(1): 77–89.

Holland S, Hosny A, Newman S, et al. (2018) The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*.

Ienca M and Malgieri G (2021) Mental data protection and the gdpr. *Available at SSRN 3840403*.

Jabbar R, Al-Khalifa K, Kharbeche M, et al. (2018) Real-time driver drowsiness detection for android application using deep neural networks techniques. *Procedia Computer Science* 130: 400–407.

Jay S (2017) What's wrong with airport face recognition? *ACLU* https://www.aclu.org/blog/privacy-technology/surveillance-tec.

Jiang M, Francis SM, Srishyla D, et al. (2019) Classifying individuals with asd through facial emotion recognition and eye-tracking. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 6063–6068.

Johanssen JO, Bernius JP and Bruegge B (2019) Toward usability problem identification based on user emotions derived from facial expressions. In: *2019 IEEE/ACM 4th International Workshop on Emotion Awareness in Software Engineering (SEmotion)*. IEEE, pp. 1–7.

Khan RA, Crenn A, Meyer A, et al. (2019) A novel database of children's spontaneous facial expressions (liris-cse). *Image and Vision Computing* 83: 61–69.

Ko BC (2018) A brief review of facial emotion recognition based on visual information. *Sensors* 18(2): 401.

Kołakowska A, Landowska A, Szwoch M, et al. (2014) Emotion recognition and its applications. In: *Human-Computer Systems Interaction: Backgrounds and Applications 3*. Springer, pp. 51–62.

Kosti R, Alvarez JM, Recasens A, et al. (2017) Emotion recognition in context. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1667–1675.

Krippendorff K (2004) Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research* 30(3): 411–433.

Krippendorff K (2018) *Content Analysis: An Introduction to Its Methodology*. Los Angeles, CA: Sage publications.

Landis JR and Koch GG (1977) An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 33(2): 363–374.

Langner O, Dotsch R, Bijlstra G, et al. (2010) Presentation and validation of the radboud faces database. *Cognition and Emotion* 24: 1377–1388.

Le Mau T, Hoemann K, Lyons SH, et al. (2021) Professional actors demonstrate variability, not stereotypical expressions, when portraying emotional states in photographs. *Nature Communications* 12(1): 1–13.

LeDoux JE and Hofmann SG (2018) The subjective experience of emotion: A fearful view. *Current Opinion in Behavioral Sciences* 19: 67–72.

Li S and Deng W (2018) Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing* 28(1): 356–370.

LoBue V and Thrasher C (2015) The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in Psychology* 5: 1532.

Mantello P, Ho MT, Nguyen MH, et al. (2021) Bosses without a heart: Socio-demographic and cross-cultural determinants of attitude toward emotional ai in the workplace. *AI & Society*. 1–23.

Mari L (2003) Epistemology of measurement. *Measurement* 34(1): 17–30.

McStay A (2020) Emotional ai, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. *Big Data & Society* 7(1): 2053951720904386.

McStay A and Urquhart L (2019) 'this time with feeling?' assessing eu data governance implications of out of home appraisal based emotional ai. *First Monday* 24(10).

Mehendale N (2020) Facial emotion recognition using convolutional neural networks (ferc). *SN Applied Sciences* 2(3): 1–8.

Mehrabian A (1996) Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology* 14(4): 261–292.

Mellouk W and Handouzi W (2020) Facial emotion recognition using deep learning: Review and insights. *Procedia Computer Science* 175: 689–694.

Meuwissen AS, Anderson JE and Zelazo PD (2017) The creation and validation of the developmental emotional faces stimulus set. *Behavior Research Methods* 49(3): 960–966.

Mollahosseini A, Hasani B and Mahoor MH (2017) Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10(1): 18–31.

Mozur P (2020) One month, 500, 000 face scans: How china is using a.i. to profile a minority. *New York Times* www.nytimes.com/2019/04/14/technology/china-surveillance-art.

Mukhopadhyay S and Sharma S (2020) Real time facial expression and emotion recognition using eigen faces, lbph and fisher algorithms. In: *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, pp. 212–220.

Munn L (2020) *Logic of Feeling: Technology's Quest to Capitalize Emotion*. Lanham, MD: Rowman & Littlefield Publishers.

Neuendorf KA (2017) *The Content Analysis Guidebook*. Los Angeles, CA: Sage publications.

Nicolaou MA, Gunes H and Pantic M (2011) Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing* 2(2): 92–105.

Nojavanasghari B, Baltrušaitis T, Hughes CE, et al. (2016) Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. pp. 137–144.

Northcutt CG, Athalye A and Mueller J (2021) Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*.

Paiva-Silva AId, Pontes MK, Aguiar JSR, et al. (2016) How do we evaluate facial emotion recognition? *Psychology & Neuroscience* 9(2): 153.

Paullada A, Raji ID, Bender EM, et al. (2021) Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2(11): 100336.

Sambasivan N and Veeraraghavan R (2022) The deskilling of domain expertise in ai development. In: *CHI Conference on Human Factors in Computing Systems*. pp. 1–14.

Stark L (2018) Facial recognition, emotion and race in animated social media. *First Monday* 23(9). doi:10.5210/fm.v23i9.9406. https://firstmonday.org/ojs/index.php/fm/article/view/9406

Stark L and Hoey J (2021) The ethics of emotion in artificial intelligence systems. In: *FACCT: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. pp. 782–793.

Stark L and Hutson J (Forthcoming) Physiognomic artificial intelligence. *Fordham Intellectual Property, Media & Entertainment Law Journal* Available at SSRN: https://ssrn.com/abstract=3927300.

Vardarlier P and Zafer C (2020) Use of artificial intelligence as business strategy in recruitment process and social perspective. In: *Digital Business Strategies in Blockchain Ecosystems*. Springer, pp. 355–373.

Vemulapalli R and Agarwala A (2019) A compact embedding for facial expression similarity. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5683–5692.

Wakefield J (2021) Ai emotion-detection software tested on uyghurs. *BBC News* https://www.bbc.com/news/technology-57101248.

Wang K, Zhu Z, Wang S, et al. (2016) A database for emotional interactions of the elderly. In: *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*. IEEE, pp. 1–6.

Wang S, Liu Z, Lv S, et al. (2010) A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia* 12(7): 682–691.

Washington P, Kalantarian H, Kent J, et al. et al (2021) Training affective computer vision models by crowdsourcing soft-target labels. *Cognitive Computation* 13(5): 1363–1373.

Wearne T, Osborne-Crowley K, Rosenberg H, et al. (2019) Emotion recognition depends on subjective emotional experience and not on facial expressivity: Evidence from traumatic brain injury. *Brain Injury* 33(1): 12–22.

White SW, Abbott L, Wieckowski AT, et al. (2018) Feasibility of automated training for facial emotion expression and recognition in autism. *Behavior Therapy* 49(6): 881–888.

Yan WJ, Li X, Wang SJ, et al. (2014) Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS One* 9(1): e86041.

Yang T, Yang Z, Xu G, et al. et al (2020) Tsinghua facial expression database–a database of facial expressions in chinese young and older women and men: Development and validation. *PloS One* 15(4): e0231304.

Zepf S, Hernandez J, Schmitt A, et al. (2020) Driver emotion recognition for intelligent vehicles: A survey. *ACM Computing Surveys (CSUR)* 53(3): 1–30.