On the Praxes and Politics of AI Speech Emotion Recognition

Edward B. Kang
University of Southern California
byungkwk@usc.edu

ABSTRACT

There is no scientific consensus on what is meant by "emotion" – researchers have examined various phenomena spanning brain modes, feelings, sensations, and cognitive structures, among others, in their study of emotional experiences. For the purposes of developing an AI speech emotion recognition (SER) system, however, emotion must be defined, bounded, and instantiated as ground truth in the training data. This means practical choices must be made in which particular emotional ontologies are prioritized over others in the construction of SER datasets. In this paper, I explore these tensions around fairness, accountability, and transparency by analyzing open-source datasets used for SER applications along with their accompanying methodology papers. Specifically, I critique the centrality of discrete emotion theory in SER applications as a contestable emotional framework that is invoked primarily for its practical utility and alignment – as opposed to scientific rigor – with machine learning epistemologies. In so doing, I also shed light on the role of the dataset creators as emotional *designers* in their attempt to produce, elicit, record, and index emotional expressions for the purposes of crafting SER training datasets. Ultimately, by further querying SER through the aperture of Critical Disability Studies, I use this empirical work to examine the sociopolitical stakes of SER as a normative and regulatory technology that siphons emotion into a broader agenda of capitalistic productivity in the context of call center optimization.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Philosophical/theoretical foundations of artificial intelligence; • **Human-centered computing** → human computer interaction (HCI); HCI theory, concepts, and models; • **Social and professional topics** → Professional topics; Computing profession.

KEYWORDS

Emotion AI, speech emotion recognition, affective computing, critical study of AI, disability and AI, social critique of AI

ACM Reference Format:
Edward B. Kang. 2023. On the Praxes and Politics of AI Speech Emotion Recognition. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3593013.3594011

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594011

1 INTRODUCTION

"Let me tell you something."

41% Fear + 14% Anger + 14% Sadness + 7% Confusion + 7% Distress + 3% Amusement + 3% Disappointment + 3% Ecstasy + 3% Embarrassment + 3% Surprise (positive) [22].

According to a study examining the ability to detect emotion from speech conducted by Cowen et al [23][1], of the eleven emotions that characterize the utterance, "Let me tell you something" by speaker 148 of the Vocal Expressions of Nineteen Emotions across Cultures (VENEC) corpus[2] [41], "Fear" is the most indicative of this speaker's emotional state. Indeed, when one listens to the vocal sample, a likely assumption might be that the speaker does not sound particularly "happy." Whether the tone is more indicative of frustration, irritation, confusion, or fear, however, is contestable. Presumably, one's perception or classification of a tone as sonically representing a particular emotion would depend on personal experiences that allowed one to form mental associations between certain tonal features of speech with particular emotional signals represented by learned emotion labels. The sound of "fear" for one person can vary significantly from that of another.

So, what "is" emotion, and how does one know, or in this case, *hear* it? To continue the example of fear, if it is an emotional construct that varies from person to person, it also means that its boundaries are not definitive but fluid. Where does fear begin, and when does it stop? Which 41% of the utterance by speaker 148 constitutes "Fear"? Does the fact that 41% of the annotators perceived the utterance to reflect "Fear" ascertain the sample's audible fearfulness? Which part of the utterance correlates with the 3% of ecstasy? Or the 14% of anger? Or the 3% of amusement? The absurdity of these questions is natural, and reflective of the immense ambivalence that characterizes what exactly one means when referring to "emotion." Indeed, as psychologist James A. Russell writes, "there are no formal criteria for what is and what is not an emotion" [62]. If anything, he suggests that the term "emotion" should represent a general banner under which different events, states, reflexes, sensations, feelings, and attitudes, among others, can be generally discussed, without necessarily having the power to denote ontological boundaries, both in the sense of categorizing particular phenomena *as* emotions, as well as in the sense of differentiating between what *is* and *isn't* an emotion.

This deconstructive move by Russell is a response to the theory of "discrete emotions[3]," which uses labels such as "fear," "anger," and

[1] Access the interactive visualization at [22].

[2] From the Proceedings of the LREC 2010 Workshop on Corpora for Research on Emotion and Affect: "The VENEC corpus consists of 100 professional actors from 5 English speaking cultures (USA, India, Kenya, Singapore, and Australia) who vocally expressed 19 different affects/emotions (affection, amusement, anger, contempt, disgust, distress, fear, guilt, happiness, interest, lust, negative surprise, neutral, positive surprise, pride, relief, sadness, serenity, and shame), each with 3 levels of emotion intensity, by enacting finding themselves in various emotion-eliciting situations." [41]

[3] The theory of discrete emotions is also referred to as a "categorical" view of emotion or simply "basic" emotion theory.

"joy," among others, to refer to basic and biologically determined states that are posited to exist across all humans. Influenced by Charles Darwin's writings in *The Expression of the Emotions in Man and Animals* [25], psychologists such as Silvan Tomkins [71, 72], and later, Paul Ekman [30], popularized the notion that prototypical and universal emotions can be traced back to evolutionarily developed behaviors and physiological reflexes[4]. Despite the popularity of the theory, however, numerous scholars have pointed out that this attempt to reliably link emotional experiences to physiological patterns was never fully realized (e.g., [11, 27, 62]). While there is indeed evidence that points to correlations between emotion, physiology, and behavior, no scientific convergence exists on the exact contours, strength, or reliability of those correlations.

To once more return to the earlier example of "fear," consider the well-known example provided by William James [75] and the bear. Here, James theorizes that the evolutionary act of physically fleeing from a predator in the wild leads to the psychological sensation of fear (emotion) – it is not the bear, but the psychological sensation of fleeing from danger that elicits the emotion. Of course, as Russell points out, this "fear" is not the same emotion that one might experience watching a horror film – there is no danger, there is no escape, and the experience is arguably enjoyable for the person paying to watch the film: "What, other than the label *fear*, do various instances of fear share with each other that they do not share with what is not fear? . . . There may be no one scientific model that applies to all cases of fear, and only to fear" [62].

I begin this paper grappling with the ontological and epistemological complexities of emotion, or more specifically, emotion *science*[5], because I understand these complexities to be crucial mediators in conceptualizing, designing, building, and deploying any technological iteration that might be housed under the banner of "Speech Emotion Recognition" (SER)[6]. To draw from Kang's [39] work on voice biometric technologies in which he questions the scientific reliability and technical efficacy of Automatic Speaker Recognition (ASR) systems by examining them through the aperture of vocal ontology, any machine learning (ML) model that is made to recognize, detect, and classify qualitative phenomena can only do so through an internal translation of that qualitative phenomena into a statistically legible object. This is referred to in ML as "ground truth[7]." In an SER system, dynamic and complex concepts such as speech, voice, and emotion must all be stabilized through such truth labels, which means that these ontologically dynamic phenomena not only need to be bounded, but also rendered epistemologically coherent: emotion is x and it can be known (or heard) through y. In this paper, I build upon the foundational work of Stark and Hoey [63] on the unstable paradigms of emotions used in "emotional artificial intelligence" or EAI [45] systems, and show that these methods of knowing, classifying, and stabilizing in machine learning SER applications also invoke particular ontologies of emotion and voice while ignoring or inadequately engaging with others, and in the process, require *redesigning* and *reproducing* a reduced 'emotion-fit-for-ML' proxy as SER ground truth training data. Furthermore, to analyze how these scientific shortcomings in the construction of ML datasets manifest in actual commercial applications, I examine the use of SER in call center optimization through the aperture of critical disability studies to parse the relations between capitalistic productivity, worker surveillance, and emotion regulation. Ultimately, I argue that SER is a technology founded on tenuous assumptions around the science of emotion that not only render it technologically deficient but also socially pernicious.

2 STRATEGIC ONTOLOGIES: FITTING EMOTION TO MACHINE LEARNING

"We are constantly developing more accurate methods of measuring human emotions . . . It is possible, then, to speak of the emotion quantitatively, as being present in large or small amounts" [27, 52].

The concept of emotion as a quantifiable and measurable object is an appealing and provocative one from the perspective of science. As historian and medical anthropologist, Otneil Dror, recounts, the numericization of emotion, or what he calls the new "*emotion-as-number*", was a powerful construct and stimulus in the 'scientification[8]' of emotion during the late nineteenth-century: ". . . specific physiological patterns signified an emotion. . . [it] was a pattern written in the language of the biological elements that one monitored in, or sampled from, the organism – translated into a sequence of numbers" [27]. This scientific bridge constructed between emotion and physiology, of course, was indicative of a broader 'body-as-machine' schema that permeated conceptions of human behavior and body at the time. Similar changes were also observable in changing ontologies and epistemologies of voice during the late 1800s, when the pseudo-science of craniometry was thought to be a newfound solution for understanding and justifying racial differences in singing capacity and vocal timbre [28]. In this way, the inclusion of concepts such as emotion and voice into the common scientific domain of physiology produced new correlatable features through which they could be described.

There were many attempts during the late nineteenth and early twentieth centuries that drew from such physiological understandings as the bases for the development of "emotion-reading" machines. Some of these included the "Emotograph" of the 1920s, which resembled "a small radio with dials and tubes and a sort of stock market tape at one end to record the emotions" [2], or the "Emotion-Meter" developed by the head of Paramount Studio's sound department, Loren L. Ryder, "to record the spectator's heartbeat and rate of breathing as scenes of love, violence and excitement unfold[ed] upon the screen" [1]. Although these inventions were evidently based on tenuous scientific links made between physiological phenomena and emotion, they were instrumental in and

[4] See Ekman [31], Izard [37], and Tomkins [71, 72] for more detailed descriptions and support for "discrete emotion theory." Crawford [24] also provides a historical overview and critique of Ekman's work in the context of facial emotion recognition technologies.

[5] It should be made clear, here, that the ontological and epistemological complexities of emotion are not necessarily as salient in everyday contexts, in which labels such as "fear" and "anger" can and are used to characterize a vast range of events, sensations, and experiences rather smoothly. They become salient in scientific contexts in which these semantic emotional labels must be empirically established to bound particular configurations of physiology and behavior as particular emotions.

[6] See Schuller [65] for a more comprehensive history of SER.

[7] Kang [40] proposes a methodological framework called "ground truth tracings" (GTT) to understand and evaluate machine learning systems by examining the processes of translation involved in ground truth construction.

[8] Drawing from the field of Science & Technology Studies (STS), I argue that concepts need to be massaged and redefined to be made legible to science. In this way, concepts such as emotion are *made* scientific.

Table 1: Summary of human vocal effects commonly associated with the labeled emotions as found by Murray and Arnott. Descriptions are in relation to what the authors call neutral speech. Table is adapted from [48].

	Anger	Happiness	Sadness	Fear	Disgust
Speech rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much slower
Pitch average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice quality	Breathy, chest tone	Breathy, blaring	Resonant	Irregular voicing	Grumbled, chest tone
Pitch changes	Abrupt, on stressed syllables	Smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflections
Articulation	Tense	Normal	Slurring	Precise	Normal

reflective of the broader transcription of emotion into the scientific language of numbers and graphs. As Dror writes, "the number was an important technology for the reframing of 'emotion' and its integration into the discourse of the laboratory" [27]. It was not until the turn of this century, however, when Dellaert et al applied pattern recognition techniques – i.e., machine learning – to "classify *utterances* according to their emotional content [emphasis added]" [26] that the premise of an emotion-*hearing* machine started to gain traction in the scientific community.

Outlined in their paper, "Recognizing Emotion in Speech," Dellaert et al "recorded a corpus containing emotional speech taken from the believable agent domain. . . of over 1000 utterances from several different speakers" [26]. It is important to understand, here, what the authors mean by "believable agent." This is a term coined by computer scientist Joseph Bates [12] in which he draws from the work of Disney animators, Thomas and Johnston [70], and their philosophies for expressing emotions through animated characters, to construct self-animating creatures or 'believable agents' called "Woggles." The three tenets that Bates took from the Disney animators were (1) emotional states must be unambiguously defined so that viewers can attribute definite emotional status to the character, (2) emotions must be directly apparent in the action of the character, and (3) emotional expression must be accentuated through foreshadowing and exaggeration even if it requires toning down other simultaneous actions because viewers cannot grasp emotional states immediately. The speakers recruited for Dellaert et al's study were given similar guidelines in the way they were asked to recite sentences accompanied by one of four emotion labels among 'happiness,' 'sadness,' 'anger,' and 'fear.' In other words, a particular *theory* of emotion, namely one that subscribes to the notion of "discrete" emotions, as well as a particular *method* of emotional expression, one that is aimed at maximizing at all costs the identifiability of a discrete emotion, were inscribed into the conceptual fabric of SER research from its very inception.

Indeed, even when Rosalind Picard [56] first mentioned the concept of computers recognizing emotional speech in her field-defining text, *Affective Computing* (published a year before Dellaert et al's first actual application of pattern recognition techniques to SER), she included a table (Table 1) taken from the Acoustical Society of America that summarized "the vocal effects most commonly associated with *five basic emotions* [emphasis added]" [48].

Despite the relative simplicity of the table, it should be noted that Murray and Arnott take great care in their original paper to engage with the different theories of emotion, including those of non-categorical models. Their table, however, is a product of an expansive literature review of existing studies around vocal features and emotion at the time. Because the majority of these studies organized emotions across the five categories labeled in the table, there is an absence of non-categorical representations of emotion. Picard's adaptation of this table as a response to the question she self-poses – "what is a good computational mapping between emotions and speech patterns?" [56] – for the purposes of speech-based affective computing is reflective of a common practice in machine learning, in which data, frameworks, and theories are imported from relevant 'external' domains as ground truths for a particular ML system. The issue with such importations, however, is that once a ground truth for a machine learning task is established and accepted, a research community forms around it through which that ground truth itself becomes the basis for the development and testing of further evaluation benchmarks and performance metrics. This further entrenches it as the norm for how that task is conceptualized and experimented with [38]. Indeed, these assumptions, coincidentally informed by a philosophy of animation (a historical detail evidently buried given the absence of mentions in subsequent SER research papers) and a turbulent research community around theories of emotion in the field of psychology, are taken as ground truth in the vast majority of highly cited SER research papers published in the decade or so that follows (e.g., [36, 53, 54, 66, 67]).

Although the influence of both Dellaert et al's study and Picard's book in adopting a discrete theory of emotion and a modular view of speech/voice as ground truth cannot be ignored, it would be inaccurate to say that their initial adoption alone explains why competing ground truths were not eventually established. Indeed, both a categorical view of emotion as well as vocal features described through dimensions such as pitch and frequency present useful taxonomies and measures within which the statistical processes of machine learning can be grounded. A categorical view of emotion, which posits that all emotional experiences can ultimately break down into around six discrete states (the number varies based on the particular theory, but the premise of separate buckets remains unchanged), presents a convenient framework of emotional taxonomy compatible with the classification logics of machine learning. In fact, this taxonomic representation and the identification of organizing labels are *required* processes in ground-truthing a qualitative phenomenon as a machine learning problem [38, 40]. There is thus a practical advantage in machine learning to maintaining a categorical view of emotion as opposed to a dimensional [62] or constructivist [11] model, both of which emphasize (to varying degrees) the fluidity of emotion as a phenomenon that is contingent on changing situations and individuals, and thus difficult to stabilize/organize as 'biologically hardwired.' The alignment between discrete emotion theory and ML emotion recognition can thus be seen as a utilitarian one, strategically prioritizing formal compatibility over scientific depth[9].

2.1 Discrete Emotion Theory vs. Core Affect

A 2019 study by Cowen et al (also referenced in the introduction), however, makes a stronger claim with regards to the *scientific* validity of discrete emotion theory in SER. Here, the authors set out to analyze "what drives the recognition of emotion, emotion categories (for example, Awe or Fear) or broader scales that capture core affect appraisals (Valence and Arousal[10])" [23]. It is important to parse their approach to this question because it specifically addresses a gray area in previous SER research studies, which have largely glossed over the lack of scientific consensus on emotional ontology and have instead relied on community-norms around categorical views of emotion as ground truth for evaluation and performance benchmarks. Indeed, machine learning domains – i.e., "emotion recognition – and the communities that form around them are generally framed around "challenges" or "problems" to be solved, which means there is a key problematization process of turning a messy qualitative phenomenon into a usable quantifiable object [40]. By explicitly framing their study around a scientific comparison of what actually "drives the recognition of emotion," in the context of SER, Cowen et al undertake an ambitious project to move beyond a *utilitarian* framing of 'emotion-as-category' to a *scientific* justification for it.

To test this, Cowen et al recruited and asked 2,345 participants from the United States and India to "judge at least 30 randomly selected speech samples from the VENEC corpus of 2,519 speech samples" [23]. The participants were randomly assigned to one of two response formats: one group was asked to select one of 30 emotional labels such as Anger, Embarrassment, Fear, and Surprise to evaluate the emotion expressed in the speech sample, while the other group was asked to evaluate speech samples based on 23 different affective features using 9-point Likert scales (see "Supplementary Table 2" in [23]). Some of the questions that the participants were asked in relation to these affective features included those inspired by componential models [64] such as "To what extent does the speaker feel like his/her situation is compatible with his/her self-image?" or "to what extent does the speaker feel like he/she can adjust to his/her circumstances?" in addition to questions pertaining to more traditional dimensional models [62] such as "to what extent does the speaker feel pleasant?" or "to what extent does the speaker feel stimulated?".

The research question and study design can be seen as a response to Russell's theory of "core affect," in which he dismantles the categorical view of emotion based on its lacking capacity to establish a consistent method and scientific ground for defining what is and isn't an emotion [62]. Highlighting that many scientific studies of emotion actually examine different phenomena spanning "brain modes, actions or action tendencies, reflexes, instincts, attitudes, cognitive structures, motives, sensations, [and] feelings" [62], Russell proposes that for the *scientific* study of emotion, which is an amorphous and composite concept, there needs to be a universal and irreducible unit to describe emotion*al* experiences. He thus proposes the core affect model, a dimensional framework that orthogonally maps Valence (positive vs. negative), Activation (stimulated vs unstimulated), and Dominance (in control vs out of control) in a two- or three-dimensional space (Valence and Activation are the primary axes, with Domination as an occasional third), which allows for various emotional experiences to be described. For instance, the emotional experience associated with the category "Fear" in the context of encountering a bear in the woods would occupy a point in the intersecting dimensions of negative valence, high activation, and low dominance. "Fear" felt in the context of a horror movie would be the same but with neutral or higher dominance. In this way, he argues that it can accommodate emotional categories without subscribing to the label as a fixed or universal ontology. It is thus important to understand that Russel's model of core affect is not necessarily meant to replace the social utility of emotional categories, but rather to provide a lower level scientifically sound ontology of affect that prioritizes transportability and irreducibility at the expense of everyday relatability. The focus is to develop *scientific tools* to describe emotional experiences writ large, of which emotional categories are a part.

This background from which the dimensional theory of core affect emerged must be taken seriously in the context of Cowen et al's study because it changes the nuances of their hypothesis that "if categories of emotion (for example, Amusement) are psychologically constructed from more basic appraisals of core affects (Valence and Arousal), one would expect the recognition of emotion in prosody along scales such as Valence and Arousal to be better preserved across cultures than the recognition of emotion categories" [23]. What Cowen et al fail to address, however, is that the core affect model was not intended as a means for the "better recognition" of emotion, and was strategically positioned against "folk theories" of emotion – i.e., the categories people use to think and talk about emotion in their everyday lives – to reconcile the scientific inconsistencies around emotional categories for the specific purpose of describing emotion within a scientific framework. Cowen et al's hypothesis thus amounts to a straw man argument that incorrectly pits core affect against emotional categories in the context of everyday emotional description. That the authors find "recognition of a number of emotion categories from prosody is better preserved across *cultures* [by which they mean individuals living either in the US or India] than that of any of the 23 affective scales that [they] considered, including Valence and Arousal [emphasis added]" [23] is thus an expected result that is largely predictable with what discrete emotion theory and the dimensional model each

[9] In a widely cited paper accompanying the RAVDESS multimodal emotion dataset, Livingstone and Russo explicitly state that "While the discrete model of emotion has been criticized. . . it is a practical choice in the creation and labeling of emotion sets" [42].

[10] Arousal is often used interchangeably with Activation in psychology literature around emotion.

stand for[11]. Not only that, but given that several of the prompts provided to the "dimensional group" also asked ostensibly ambiguous or irrelevant questions such as "To what extent does the speaker feel like his/her situation is compatible with his/her self-image?", which would inevitably require participants to have acquired supplementary information around a speaker's "self-image" based on one utterance, it is unsurprising to find that judgments made by the evaluators in the "categorical group" were more consistent.

In this way, despite their ambitious attempt, Cowen et al's study falls short both methodologically and theoretically in establishing the grounds for discrete emotion theory as a more robust scientific framework to other models such as core affect in the context of SER. More accurately, it represents a fieldwide attempt to retroactively bound emotional ontology within the epistemological constraints of machine learning – i.e., to start with machine learning as the logical foundation through which qualitative phenomena, such as emotional speech, are understood, as opposed to starting with a qualitative phenomenon and finding a method (machine learning or not) that engages with it both holistically and accurately. This is a strategic reversal from the perspective of machine learning practitioners because if ML is understood as the base logic through which "external" problems can be solved, nothing *can't* be distilled into a machine learning problem. Once a qualitative phenomenon is turned into a machine learning problem, however, it should also be understood that it effectively exists as a separate "thing" tied to the epistemologies of machine learning benchmarks, techniques, and models that define it. In the next section, I turn my attention to the processes that create this separate proxy – i.e., to the actual practices of ground-truthing vocalized emotion for the purposes of building emotional speech corpora – to examine not only how emotion is conceptualized in SER, but also how it is elicited, recorded, indexed, and contained in SER training datasets.

3 GUIDED DESIGN: PRODUCING EMOTIONS FOR SER DATASETS

Momentarily assuming that the emotional ontologies invoked in SER provide apt scientific models for classifying emotional speech, it is still important to trace how those ontologies are instantiated and realized as ground truth in the training data. There must be instances of emotion captured and aggregated as data upon which an SER model can train. Simply put, it is important to look at exactly *how* SER datasets are constructed. Here, I examined 10 highly circulated and widely cited open-source datasets used in SER applications and their accompanying research papers. Given the notorious inaccessibility of commercially developed SER applications and their underlying datasets [16], I determined that these open-source alternatives would serve as generally representative samples of how SER datasets are constructed across both academia and the private tech industry based on research that shows open-source datasets also inform the construction of comparable privately developed commercial ones [9]. Some of the emotion datasets I examined include the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP), a corpus that contains audiovisual data related to performances of "selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions (happiness, anger, sadness, frustration, and neutral state)" [18], and the Montreal Affective Voices (MAV), a dataset comprising non-verbal affect bursts and explicitly framed as an "auditory counterpart of the Ekman faces" [13], among others. It should be noted that although some of the databases I engaged with were multimodal (e.g., [10, 18, 42]), I still included these corpora in my analysis due to a combination of their primary focus on vocal expressions of emotions (over facial) evident in the original papers and actual datasets, their frequent usage in SER applications, and their high number of citations. See Table 2 for list of examined SER datasets.

A common pattern across all of the original papers accompanying the major datasets I examined was the use of actors to produce emotional expressions. Indeed, as Bänziger, Mortillaro, and Scherer write, "the procedure for gathering [emotional expressions] is probably the most sensitive choice to be made, one that is often determined by the theoretical assumptions of the researchers" [10]. Broadly, these procedures follow two primary methods of emotional production: "Communication Effect Acting (CEA)," which "requires that an expresser produces an emotion, generally specified by an emotional label — most often a label that is associated with a discrete emotion category — in such a way as to optimize the recognition of the emotion by external observers" [10], or "Felt Experience Acting (FEA)", which uses "vivid mental imagery techniques and specifically [invites] expressers to recall and relive personal past events when they felt the target emotion" [10]. The procedure for "gathering" emotional expressions thus always requires a process of "producing" them first.

3.1 Communication Effect Acting (CEA)

Consider, for instance, the aforementioned "believable agent domain" used by Dellaert et al in their seminal 1996 study – in order to test "statistical pattern recognition techniques to classify utterances according to their emotional content" [26], they needed to first *record* a corpus of emotional speech *produced* by actors who were given *scripts* with emotional *labels* and instructed to *exaggerate* their emotional expressions to *maximize* perceptibility. Dellaert et al were undeniably a part of the production of the emotions they were measuring. This was also the case in the construction of the Montreal Affective Voices (MAV), the self-proclaimed auditory counterpart to the Ekman faces:

> The actors were instructed to produce short emotional interjections, using the French vowel *ah*. . . and were played an auditory demonstration of the expressions that they would be asked to generate before the recording session. They had to produce vocal expressions corresponding to happiness, sadness, fear, anger, pleasure, pain, surprise, and disgust, as well as a neutral expression. Each category of vocalizations was *performed several times until our qualitative criterion was reached – that is, until the affective vocalization produced was clearly recognizable by the experimenter*

[11] A comparable (but not identical) scientific tool is the Kelvin, which is a scientific unit for measuring temperature, but does not necessarily stand in to replace the experience of climate in everyday discourse over more commonly used terms such as "chilly" or "warm". For instance, one would expect the results of a study that asked participants unfamiliar with K units to describe the climate of a room via temperature-related labels vs numerically in Kelvins to show that participants are generally more likely to consistently describe a 10 degrees Celsius room as "chilly" than as 283 K.

Table 2: List of examined SER datasets.

Reference	Name	Description	Method of Emotional Production
Bänziger et al [10]	GEMEP	7000 audio-visual emotion portrayals; 18 emotions represented; portrayed by 10 professional actors coached by a professional director	Felt Experience Acting (FEA)
Belin et al [13]	MAV	90 nonverbal affect bursts; 8 emotions represented (and neutral baseline); recorded by 10 actors (5 male, 5 female)	Communication Effect Acting (CEA)
Busso et al [18]	IEMOCAP	151 videos of recorded dialogues including 2 speakers per video; 8 emotions represented (and neutral baseline)	FEA
Busso et al [19]	MSP-IMPROV	7818 non-read (improvised or natural) speech and 620 read sentences; 4 basic emotions along with "other" option are provided as initial single-choice question, along with an additional 6 emotional categories as multi-choice question; recorded by 12 speakers (6 males, 6 females)	CEA & FEA
Cao et al [21]	CREMA-D	7442 audio-visual emotion portrayals; 5 emotions along with "no emotion" option represented; portrayed by 91 actors (48 male, 43 female) coached by 1 of 2 different directors	CEA
Laukka et al [41]	VENEC	6500 audio emotion portrayals; 19 different emotions represented; portrayed by 100 professional actors	CEA & FEA
Livingstone & Russo [42]	RAVDESS	7356 audio-visual emotion portrayals of speech and song; 6 emotions along with "neutral/calm" state represented; recorded by 24 speakers (12 males, 12 females)	CEA & FEA
Martin et al [43]	eNTERFACE'05	1166 video sequences of emotional portrayals; 6 emotions along with "neutral" state represented; recorded by 42 speakers (17 discarded), 25 speakers retained in corpus	FEA
McKeown et al [44]	SEMAINE	190 video sequences of interactions between a human and an operator playing a character with 4 emotionally-charged personalities; 20 participants	FEA
Poria et al [57]	MELD	1433 dialogues taken from television show *Friends* (13,000 utterances); 6 emotions along with "neutral" and "non-neutral" states represented	FEA

> *as the one they were asked to produce. . . Constant feedback was given to the participants during the entire session so they could improve their performance* [emphasis added]. [13]

This means that an SER system trained on the MAV dataset takes a collection of emotional vocal bursts meticulously guided and crafted by the experimenters as its ground truth for voiced emotion – the dataset creators are effectively emotional *designers.* This immediately raises questions around generalizability and ethics. From a scientific perspective, how generalizable is a particular group's notion of an emotional expression beyond that context of expression? And further from a sociopolitical standpoint, how does it normalize particularly narrow and ableist notions of emotional expression? Because a necessary pre-step for "hearing" and "recording" emotion in SER is to "define" and "produce" it, the process inherently becomes a normative one that is embedded with assumptions held by both producer and recorder. As sound studies scholar, Nina Eidsheim argues, "the assumption that we can know sound, and that the meaning we infer from it is stable (and indeed essential), allows for the *projection of beliefs about people onto the sound*" [29]. By "listening to listening" – i.e., examining *how* one listens, or in this case, examining *how* an SER application must create the sounds it needs to listen for – it becomes possible to recognize that the self-proclaimed "universal" emotions collected for MAV using CEA are actually a group of narrowly defined and guided emotional *performances*.

3.2 Felt Experience Acting (FEA)

Other datasets such as the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [18] and the GEneva Multimodal Emotion Portrayals (GEMEP) [10] try to circumvent the perceived "inauthenticity" of performed emotional expressions by instructing actors to subscribe to the aforementioned "Felt Experience Acting" (FEA) technique. This includes the use of hypothetical scenarios, interactions, monologues, and improvisations to "increase the authenticity" of the emotional expression so as to argue that the emotion is *elicited* as opposed to *performed*. Considering that such a procedure undeniably qualifies as human-subjects research, however, which requires study facilitators to explain in their Institutional Review Board (IRB) protocol how they intend to protect participants from experiencing any kind of psychological discomfort or harm, it is unclear how emotional experiences that *require* psychological discomfort such as fear, anxiety, or sadness are ethically elicited. There are no mentions of how this conundrum was addressed in any of the original papers accompanying the major datasets I examined. In such absence, I am led to deduce that in order for the elicitation to have been granted IRB approval, the elicitation process ultimately had to be redefined as one in which the emotional expressions related to psychological discomfort were not really "felt," or felt in a way that did not fully capture what one may consider the psychological discomforts related to disconcerting emotional experiences. This puts into question the efficacy of FEA in addressing the methodological limitations of CEA.

In relation to this, Bänziger et al dedicate a section in their paper accompanying GEMEP titled "In Defense of Felt Experience Enacted Emotion Expressions" [10]. Here, they accurately elaborate on the precarity and mediated nature of "naturally occurring EEs [emotional expressions] captured in everyday life" stating that there are "extraordinary ethical and practical difficulties of recording private, largely unpredictable, and fleeting episodes for a sufficient number of individuals" as well as methodological difficulties related to "sampling a large number of different emotions per person that are comparable in nature and context across individuals" [10]. Not only this, but the authors themselves also emphasize the inherently mediated nature of emotion referencing scientific literature that shows "emotional expression is rarely free of posing" [10]. Critiquing studies which used emotional expressions recorded from television shows as ground truth for 'emotions-in-the-wild' (e.g., [57]) the authors state that this mediation effect is "even worse in EEs [emotional expressions] recorded form the media (as in reality or game shows), given that in addition to ordinary display rules and regulation attempts, there may be systematic bias resulting from coaching of participants by the organizers or audience reaction" [10].

What is astonishing about this thorough description, however, is its juxtaposition against a purely utilitarian justification for FEA. While the recruiting of professional actors and recording them in controlled settings as part of FEA does allow for systematic experimentation, it does not circumvent any of the complications related to the stability, generalizability, and authenticity of emotional expression that the authors themselves critique. Even if one accepts the emotional authenticity of an actor's "attempt to enact as faithfully as possible the recalled or imagined emotional experience in the expression shown" [10], this does not address the fact that the production of emotion using FEA always creates an *instance* of emotion tied to the context of production. Put differently, a performed emotion is still an emotional experience, but it is not an emotional experience that is generalizable beyond that instance. A woman reading a sentence accompanied by the label "Joy" in a way that both she and the study facilitators perceive to be joyful may indeed be an instance of joy, but it does not represent a generalizable (neither across- nor within-subject) representation of joy. This is true regardless of whether the emotional expression is the product of a script or an "authentic" elicitation. An adequate engagement with the multiplicity of emotional ontology – which would require acknowledging that a universal experience of emotion simply does not exist and that the attempt to do so actually materializes as a form of emotional regulation that normalizes particular emotional experiences over others – shows that the limitations of SER manifest not only in the process of ground-truthing, but also in the underlying premise itself[12]. Indeed, drawing from emotion literature in psychology, existing critical scholarship on AI emotion recognition has accurately pointed out that "external" expressions of emotion comprise only a part of the broader emotional experience, and are often actually suggestive of social motivations, while providing little evidence of "interior" mental states [16, 63]. Finally, the unavoidable effect of culture, broadly construed, as a powerful mediator in the way emotions are expressed and perceived [33, 61] further sheds light on the limitations of a generalizable emotion recognition system.

Examining both the theoretical framework through which emotion is defined (discrete emotion theory) as well as the methodological choices for producing emotion at the level of ground truth (CEA and FEA) sheds light on the weight of utilitarian fit as a mode of decision-making in conceptualizing and building machine learning SER systems. This prioritization of functional proxies over scientific

[12] In the context of facial emotion recognition (FER), Cabitza, Campagner and Mattioli show that there is an "unbearably" low reliability for FER ground truths, which means that there is extreme variability in people's emotion ratings of faces. Based on the results of their experiment, the authors assert 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" [20].

depth in machine learning is not unique to SER, but it does have particularly concerning implications as it relates to SER's use cases. In the next section, I will examine some of these existing applications through the aperture of disability, problematizing SER's manifestation as a regulatory surveillance technology that applies emotion recognition to the context of optimizing labor productivity in American call centers.

4 CONSTRUCTING DISABILITY: INTERFACE FOR PATHOLOGY AND PRODUCTIVITY

Scholars working at the intersection of disability studies and technology (e.g., [7, 46, 47, 50, 74]) have long pointed out that the "concept of 'normal,' as well as the tools and techniques for enforcing normalcy, have historically constructed the disabled body and mind as deviant and problematic" [74]. In the historical development of AI, however, which operates under the epistemological domain of statistics, and therefore necessitates normative models in their successful operation, this understanding of disability has served as a "charismatic" use case [8] for the development of "assistive" technologies[13], as well as been unsettlingly used as a metaphor for the deficient computer.

"*Helping Autistic People"* is one of the section titles and listed applications for affective computing in Picard's 1995 text [55]. Here, she states that "one of the hallmarks of [autism] is difficulty with emotions – recognizing the meanings of other peoples' emotions, suitably expressing emotions, and having empathy" [56]. She goes on to elaborate that

> . . . computers are like autistic people – particularly like autistic "idiot savants," an unfortunate term that has been used to describe people who have unusually gifted abilities in certain areas – such as rapid computation of large numbers, memorizing phone listings, and precise memory of huge sets of facts and trivia, but who lack the forms of common sense and emotional intelligence that most people acquire effortlessly. . . Unaffective computers are similarly *handicapped* [emphasis added]. [56]

Leveraging both a metaphorical alignment between autism and computation, as well as presenting automated emotion recognition as an educational tool and "assistive" technology for autistic individuals, Picard casted emotion recognition in a prosocial light, allowing her to garner grant support from the National Science Foundation, the National Institutes of Health, and the National Institute of Mental Health [50]. The problem with this metaphor, understood as either the computational model of autism – i.e., likening autistic individuals to "computers running a poorly implemented software architecture, one that, insofar as empathy is concerned, suffer[s] from an internal signal processing disorder that ha[s] concomitant external effects on social signal processing" [50] – or the autistic model of computers – i.e., 'handicapped' machines that fail to interpret and respond to users' emotions – is that it adopts a medical model of disability that understands autism as a *deficit* of the individual. Relying on biomedical standards of "normal" bodies, such a view pathologizes conditions outside of these norms as impaired and thus in need of augmentation or correction [68]. This is in opposition to a social[14] model of disability, which understands it as a *relational* product contingent on environments and attitudes, wherein "the locus of intervention [is] not at the level of the individual – with efforts that would attempt to "fix" disabled people – but at the level of social justice" [74].

Conceptualized and developed through its medical framing as a "socio-emotional prosthetic" for autistic individuals [50], AI-emotion recognition (AI-ER) thus effectively abstracts autism away from the diversely experienced condition that mediates autistic individuals' lives, into a standardized pathology that can be used to denote empathy impairment in both humans and machines. In recounting the development of El Kaliouby and Robinson's "emotional hearing aid," a facial affect prosthetic for children with Asperger's syndrome [32], Nagy observes that their adoption of a medical model of autism "prompted the system's users to shape their emotional expressions to meet the expressions of non-autistic social others" [50]. Based on real-time video assessment of detected faces, the emotional hearing aid instructed users to "apologize or explain for what the system register[ed] as a neurotypical interlocutor's disgust or confusion [which] place[d] the burden of rectifying an apparent 'conversational impediment' in the interests of producing a 'productive state' entirely on side of the autistic individual" [50]. In this way, it served as a disciplinary technology that trained autistic users towards "becoming docile and conciliatory conversation partners" [50] to their non-autistic counterparts.

As Nagy documents, this historical relationship between affective computing and autism research is an instance of what Mills has called "assistive pretext" [46], wherein technologies initially developed to address disability-related use cases are reconfigured and made relevant to the general public. The translation from socio-emotional prosthetic to commercial emotion AI, which locates emotion recognition within the extractive domain of surveillance capitalism [76], sheds light on a

> . . . productive parallelism between the ways that autism mitigation programs view autistic individuals and the ways that surveillance capitalism views platform users. Both view the subjects in their purview as collections of manipulatable, infra-individual behaviors instead of as independent actors capable of psychological depth. [50]

Disability, understood as pathology, and medical models of disability 'correction' thus comprise the sociotechnical interface through which commercial AI emotion recognition interprets and ultimately regulates its subjects. It constructs a system in which subjects are required to make their emotional experiences legible to the systems that detect it, thereby flipping the directionality of power from one that, in its most ideal iteration, should be able to "recognize" emotional expressions in all of their multiplicity to one that, in practice, "constrains" them to a set of normative behaviors.

[13] This is a shorthand used to refer to technologies that assist disabled people. It has been critiqued, however, for its redundancy – all technology is assistive, not just technology that assists disabled individuals – and paternalism. Mills has also argued that "the phrase advances a technological fix that is unconcerned with education, community support, or social change" [46].

[14] It should be noted that although the social model is useful for positioning disability beyond the frame of individual pathology, it has also been critiqued by disability scholars for leaving "impairment unchecked, undertheorized, and ignored" [34].

In the following section, I will show that this "reverse Turing-test" situation [51], in which humans must validate their humanity to AI systems, is also highly representative of the way that SER is currently deployed in commercial contexts.

4.1 Disabl*ing* Technology: SER in Call Center Optimization

Cogito is a software company that applies SER to the context of call center optimization. In monitoring conversations between call-center operators and customers, the company "extracts and analyzes over 200 acoustic and voice signals in milliseconds to give . . . agents cues on how to adjust their behavior and surface the best recommendations" [3]. It is promoted as a human-centered AI tool that improves customer experience (CX), in which "CX measured on every call delivers the data for personalized coaching and development plans, essential to growing and retaining your employees" [4]. According to its website, Cogito is used by "3 of the top 5 US telecom technology companies," "2 of the top 5 US cable providers," "4 of the top 5 national health insurers," and "3 of the top 5 pharmacy benefits managers," of which 8 are Fortune 25 brands [4]. Cogito's products thus span multiple industries and have largely become an embedded part of the CX in American call centers.

There are important parallels to consider in the way Cogito frames its use case as helping call center operators through SER with the aforementioned histories of affective computing and autism. A 2018 article in *Wired* referenced Cogito Chief Operating Officer (COO) Tracy Dudek stating that positive emotional evaluations of agents by its SER system can boost agents' chances of performance-related bonuses [69]. The flipside of this, of course, is a performance-related *penalty* related to behaviors that might be categorized as a "negative emotional evaluation[15]," a feature also observed across other AI emotion recognition systems used for worker evaluations described in various different patent applications [16]. Dudek's statement reveals an explicit link between emotional regulation and worker productivity facilitated through a surveillance tool that constantly monitors operators' conversations, in which operators are seen through a dehumanizing aperture akin to what Nagy describes in the context of autism mitigation programs [50]. Understood as a modular assemblage of adjustable behaviors as opposed to independent individuals with the capacity to experience a range of emotions, Cogito's SER effectively serves as a disciplinary mechanism that systematically fastens worker compensation to emotional regulation so as to enforce workers to self-regulate their behaviors to be positively read by Cogito's SER system. Roemmich et al have referred to this mechanism as the "*affective commodification* of a candidate's *affective value*" [58] to denote the role of AI emotion recognition in determining emotional measures of desirability for subjected workers. As I delineate in previous sections, however, an SER system, like any other ML technology, is highly contingent on its ground truth, wherein the ground truth does not necessarily denote any kind of real qualitative truth, but rather reflects the ontological alignments and epistemological assumptions of its creators. In other words, the ground truth for "positive evaluations of emotion" in Cogito's SER system is a construct designed by Cogito itself. The company's rhetoric of human-centered design and improving self-satisfaction of agents through empathy coaching can thus be seen as denoting an arguably unproductive self-contained loop in which operators are simply making themselves legible to the data structures now entwined in their performance reviews, and as an extension, their livelihoods.

There are also important practical implications of applying SER to CX on the side of influencing customer behavior. In the "Ubiquitous Technologies for Emotion Recognition" special issue for *Applied Sciences*, Bojanic, Delic and Karpov propose a system in which call centers can rank the urgency of calls based on SER, "giving greater priority to calls featuring emotions such as fear, anger, and sadness, and less priority to calls featuring neutral speech and happiness" [15]. Juxtaposing research that documents the immense psychological strains of abusive behavior by "angry" callers on call-center employees (e.g., [58, 59]) alongside the explicit use case invoked by companies such as Cogito – to "Reduce Employee Churn" [5] – a proposition to prioritize calls that are audibly "angrier" will presumably have an even greater negative effect on the mental health of call center operators given the obvious message it sends to customers with regards to "gaming" the system: sound stereotypically angrier. Not only that, but by focusing specifically on the "empathy coaching" of agents, thus disciplining them to better regulate both their own and callers' emotions, while possibly also implicitly encouraging more abusive behavior on the part of callers via the emotional ranking system, the application of SER to call centers ultimately strips power away from the operators. This is similar to how El Kaliouby and Robinson's emotional hearing aid was sold as a tool to help autistic individuals, but in reality, resulted in assimilating them to ableist standards of conversational productivity. By prompting autistic users to apologize for or rectify neurodivergent behaviors, the emotional hearing aid placed neurotypical ones as the singular mode of legible – i.e., acceptable – emotional expression, without ever fully engaging with the possibility of a wider spectrum of emotional productivity.

Indeed, as a concept that was *invented* in the nineteenth century directly alongside the notion of the wage economy [60], "when employers began to use preemployment screenings to eliminate people deemed inefficient, nonproductive, and likely to require extra help and support" [74], disability cannot be understood separately from the notion of worker productivity. In this history, the lack of ability to conform to particular notions of productivity effectively rendered an individual "disabled," thereby disqualifying the person from participating in the workplace and earning a living. Cogito's use of SER as a mode of worker surveillance continues this history of ableist workplace politics in which its subjects are problematically located on an expanded spectrum of emotional suspects that require constant monitoring and validating. As part of a scientific lineage that is intimately linked with adopting a medical – i.e., deficit – model of autism, and then applying that model as the default framework for understanding its users beyond the original context of disability, Cogito's SER can also be interpreted

[15] An FER system developed by the recruiting-technology firm HireVue similarly analyzed facial movements to denote the potential productivity and overall "employability" of a candidate [35]. HireVue eventually removed this function from their recruiting technologies not only "due to the prolonged criticism that Ekman's universality thesis has faced by cultural anthropologists and others" [17] but also in part because of the inherently unethical nature of using AI-assisted physiognomy to evaluate potential workers.

as a disabl*ing* technology that transposes the systemic oppression of disabled individuals on to its users. To be clear, this is by no means to argue that the systemic oppression of disabled people should be contained, but rather the opposite. It is to problematize the medical model of disability altogether so as to argue that such ways of thinking ultimately make such oppressive structures more widespread.

5 CONCLUSION

My analytical approach to AI speech emotion recognition in this paper can be considered twofold, in which I begin with an empirical engagement with machine learning practices and methodologies for constructing datasets, and end with a critical analysis of SER and its application to call-center optimization seen through the aperture of disability as social construction. Despite the difference in tone and analysis, the impetus for each is united in my broader hope of bringing attention to the technology's limitations and potentials for harm. As a premise that is largely dependent on siphoning contested scientific research on emotional ontology into a constrained epistemological framework of machine learning, emotion recognition via artificial intelligence still amounts to more of an aspiration for the AI community than a reality. Furthermore, given these epistemic limits of machine learning, in which the knowledge produced from an ML model is inextricably tied to its ground truth reference, the role of the dataset creator in designing emotion as part of the ground-truthing process and the "unbearable" unreliability of peoples' abilities to agree on emotional signals [20] are unavoidable. As a result, I argue that a quasi-functional data proxy for emotion is created that fails to encapsulate neither the ontological breadth nor depth of emotional experiences, but nonetheless inherits the cultural significance of emotion through its manifestation in SER.

Applying this to the context of Cogito's SER system, I further show that such data proxies and the practices through which they're constructed produce a disciplinary system that divests power from call center agents and consolidates it within its creators. This alone is not a novel critique of the power relations that mediate technology creators and users – indeed my own situating of Cogito within a longer history of affective computing research demonstrates the normative functions of emotion-recognition technologies that have come long before Cogito's SER. Instead, my intervention lies in the reinforcement of this critique through my empirical engagement with the grounded practices, techniques, methodologies, and decisions that shed light on exactly *how* these proxies are constructed, and what they show about the tenuousness of the promises of these technologies. My hope is that this intervention is taken as both a broader push towards interdisciplinary collaborations between "technical" and "sociotechnical" communities[16] (for lack of a better semantic distinction), as well as a more focused and urgent call to reevaluate the limitations underlying the premise of SER.

Although call centers currently represent the most widely observed commercial use case of SER led by technology companies such as Cogito, there is a wide variety of other emerging applications such as Amazon's emotion tracking wearable device, "Halo," which monitors and analyzes a user's tone of voice to make wearers more aware of how they sound to others [49], or Voicesense's suite of voice analytics products, one of which apparently measures emotional states from the voice to screen and monitor mental health risks, such as depression [73]. Affectiva (now a subsidiary of Smart Eye), the facial emotion recognition (FER) company founded by none other than El Kaliouby and Picard, also wrote of the importance of SER in developing more robust multimodal emotion AI systems moving forward [6]. Given the comparative slowdown in FER development with Microsoft's recent decision to retire its emotion recognition technologies from its Azure Face facial recognition services [14], it will be important to also bring attention to the dubious scientific and ethical implications of SER before it becomes a more ubiquitously adopted technology.

ACKNOWLEDGMENTS

I am grateful for the conversations I had with friends, colleagues, and mentors in the writing and editing of this paper. These include, but are not limited to: Kate Crawford, Larry Gross, Josh Kun, Mike Ananny, John Cheney-Lippold, Amy Lee, and fellow members of the Sloan-funded research collective "Knowing Machines." I also very much thank the three anonymous reviewers and program chairs of *FAccT'23* for their feedback.

REFERENCES

[1] Emotion-Meter for Use at Film Previews. Attached to letter Leonarde Keeler, December 5, 1946. Drawr/Unmarked/Misc. 4. Department of Defense Polygraph Institute.

[2] The Hidden Truth. Background Material, Joliet State Penitentiary, folder/Source Materials/I/B, Department of Defense Polygraph Institute.

[3] Cogito. Retrieved December 10, 2022 from https://cogitocorp.com/

[4] Emotion AI. Cogito. Retrieved December 10, 2022 from https://cogitocorp.com/emotion-ai/

[5] Reduce Employee Churn. Cogito. Retrieved December 10, 2022 from https://cogitocorp.com/solutions/reduce-employee-churn/

[6] Affectiva. 2017. Introducing Affectiva's Emotion Recognition through Speech. Affectiva. Retrieved August 10, 2022 from https://blog.affectiva.com/introducing-affectivas-emotion-recognition-through-speech

[7] Meryl Alper. 2018. Can Technology Really 'Give Voice' to Disabled People?. Pacific Standard. Retrieved December 15, 2022 from https://psmag.com/social-justice/can-technology-really-give-voice-to-disabled-people

[8] Morgan Ames. 2019. The Charisma Machine: The Life, Death, and Legacy of One Laptop per Child. MIT Press, Cambridge, MA.

[9] Andy Baio. 2022. AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability. (September 30). Waxy. Retrieved December 5, 2022 from https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/

[10] Tanja Bänziger, Marcello Mortillaro, and Klaus R Scherer. 2012. Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. Emotion 12, 5 (Oct. 2012), 1161-1179. DOI: https://doi.org/10.1037/a0025827

[11] Lisa Feldman Barrett. 2017. The theory of constructed emotion: An active inference account of interoception and categorization. Social cognitive and affective neuroscience 12, 11 (Jan. 2017), 1-23. DOI: https://doi.org/10.1093/scan/nsw154

[12] Joseph Bates. 1994. The role of emotion in believable agents. Communications of the ACM 37, 7 (Jul. 1994), 122-125. DOI: https://doi.org/10.1145/176789.176803

[13] Pascal Belin, Sarah Fillion-Bilodeau, and Frédéric Gosselin. 2008. The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing. Behavior research methods 40, 2 (May 2008), 531-539. DOI: https://doi.org/10.3758/BRM.40.2.531

[14] Sarah Bird. 2022. Responsible AI investments and safeguards for facial recognition. Microsoft. Retrieved December 15, 2022 from https://azure.microsoft.com/en-us/blog/responsible-ai-investments-and-safeguards-for-facial-recognition/

[15] Milana Bojanic, Vlado Delic, and Alexey Karpov. 2020. Call redistribution for a call center based on speech emotion recognition. *Applied sciences* 10, 13 (Jul. 2020), 4653. DOI: https://doi.org/10.3390/app10134653

[16] Karen Boyd and Nazanin Andalibi. 2023. Automated emotion recognition in the workplace: How proposed technologies reveal potential futures of work. In *ACM Conference on Computer Supported Cooperative Work (CSCW '23)*, October 13, 2023, Minneapolis, Minnesota. ACM., New York,

[16] See Kang [40] for an extended discussion of this.

New York, forthcoming. https://www.nazaninandalibi.net/_files/ugd/63b293_530e3aa75f034548a43140e8bda9426e.pdf

[17] Taina Bucher. Facing AI: Conceptualizing 'fAIce communication' as the modus operandi of facial recognition systems. *Media, Culture & Society* 44, 4 (May 2022), 638-654. DOI: https://doi.org/10.1177/01634437211036975

[18] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (Nov. 2008), 335-359. DOI: https://doi.org/10.1007/S10579-008-9076-6

[19] Carlos Busso, Srinivas Parthasarathy, Alex Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. In *IEEE Transactions on Affective Computing* 8, 1 (Jan. 2016), 67-80. DOI: https://doi.org/10.1109/TAFFC.2016.2515617

[20] Federico Cabitza, Andrea Campagner, Martina Mattioli. 2022. The unbearable (technical) unreliability of automated facial emotion recognition. *Big data & society* 9, 2 (Oct. 2022). DOI: https://doi.org/10.1177/20539517221129549

[21] Houwei Cao, David G. Cooper, Micahel K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sources emotional multimodal actors dataset. In *IEEE Transactions on Affective Computing* 5, 4 (Jul. 2014), 377-390. DOI: https://doi.org/10.1109/TAFFC.2014.2336244

[22] Alan S Cowen, Petri Laukka, Hilary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The emotions conveyed in speech prosody. Retrieved from https://s3-us-west-1.amazonaws.com/venec/map.html#modal

[23] Alan S Cowen, Petri Laukka, Hilary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour* 3, 4 (Apr. 2019), 369-382. DOI: https://doi.org/10.1038/s41562-019-0533-6

[24] Kate Crawford. 2021. The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven, CT.

[25] Charles Darwin. 1872. *The Expression of Emotions in Animals and Man.* Murray, London, UK.

[26] Frank Dellaert, Thomas Polzin and Alex Waibel. 1996. Recognizing emotion in speech. In *Proceeding of Fourth International Conference on Spoken Language Processing (ICSLP '96),* October 03 – 06, 1996, Philadelphia, Pennsylvania. IEEE., New York, New York, 1970-1973. https://doi.org/https://ieeexplore.ieee.org/document/606911

[27] Otniel Dror. 2001. Counting the affects: Discoursing in numbers. *Social research* 68, 2, 357-378.

[28] Nina Sun Eidsheim. 2015. Race and the aesthetics of vocal timbre. In Olivia Ashley Bloechl, Melanie Diane Lowe, & Jeffrey Kallberg, eds. *Rethinking difference in music scholarship*, 338-365. Cambridge University Press. Cambridge, UK.

[29] Nina Sun Eidsheim. 2018. *The race of sound: Listening, timbre, and vocality in African American music.* Duke University Press, Durham, NC.

[30] Paul Ekman. 1970. Universal facial expressions of emotions. *California Mental Health Research Digest* 8, 4, 151-158.

[31] Paul Ekman. 1994. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychological Bulletin* 115, 2 (Mar. 1994), 268-287. DOI: https://doi.org/10.1037/0033-2909.115.2.268

[32] Rana El Kaliouby and Peter Robinson. 2005. The emotional hearing aid: An assistive tool for children with Asperger syndrome. *Universal access in the information society* 4 (Aug. 2005), 121-134. DOI: https://doi.org/10.1007/s10209-005-0119-0

[33] Maria Gendron, Debi Roberson, and Lisa Feldman Barrett. 2015. Cultural variation in emotion perception is real: A response to Sauter, Eisner, Ekman, and Scott. *Psychological Science* 26, 3 (Mar. 2015), 357-359. DOI: https://doi.org/10.1177/0956797614566659

[34] Dan Goodley. 2018. Understanding disability: Biopsychology, biopolitics, and in an in-between-all politics. *Adapted Physical Activity Quarterly* 35, 3 (Jul. 2018), 308-319. DOI: https://doi.org/10.1123/apaq.2017-0092

[35] Drew Harwell. 2019. A face-scanning algorithm increasingly decides whether you deserve the job. (November 2019). Retrieved January 7, 2023 from https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/

[36] Hao Hu, Ming-Xing Xu, and Wei Wu. 2007. GMM supervector based SVM with spectral features for speech emotion recognition. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, April 15 – 20, 2007, Honolulu, Hawaii. IEEE., New York, New York, IV-413-IV-416. https://doi.org/10.1109/ICASSP.2007.366592

[37] Carroll E. Izard. 1994. Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin* 115, 2 (Mar. 1994), 288-299. DOI: https://doi.org/10.1037/0033-2909.115.2.288

[38] Florian Jaton. 2021. The Constitution of Algorithms: Ground-Truthing, Programming, Formulating. MIT Press, Cambridge, MA.

[39] Edward B. Kang. 2022. Biometric imaginaries: Formulating voice, body, identity to data. *Social studies of science* 54, 4 (Aug. 2022), 581-602. DOI: https://doi.org/10.1177/03063127221079599

[40] Edward B. Kang. 2023. Ground truth tracings (GTT): On the epistemic limits of machine learning. *Big data & society* 10, 1 (Jan. 2023). DOI: https://doi.org/10.1177/205395172211461

[41] Petri Laukka, Hilary Anger Elfenbein, Wanda Chui, Nutankumar S. Thingujam, Frederick K. Iraki, Thomas Rockstuhl, and Jean Althoff. 2010. Presenting the VENEC corpus: Development of a cross-cultural corpus of vocal emotion expressions and a novel method of annotating emotion appraisals. In *Proceedings of the LREC 2010 Workshop on Corpora for Research on Emotion and Affect,* Valetta, Malta. European Language Resources Association., Paris, France, 53-57. Retrieved from http://su.diva-portal.org/smash/record.jsf?pid$=$diva2%3A373848&dswid$=$7752

[42] Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13, 5 (May 2018), e0196391. DOI: https://doi.org/10.1371/journal.pone.0196391

[43] Oliver J. F. Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. 2006. The eNTERFACE'05 audio-visual emotion database. In *22^{nd} International Conference on Data Engineering Workshops (ICDEW '06)*, April 03 – 07, 2006, Atlanta, GA. IEEE., New York, New York, 8-8. https://doi.org/10.1109/TAFFC.2014.2336245

[44] Gary McKeown, Michel F. Valstar, Roderick Cowie, and Maja Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo (ICMO '10)*, July 19 – 23, 2010, Singapore. IEEE., New York, New York, 1079-1084. https://doi.org/10.1109/ICME.2010.5583006

[45] Andrew McStay. 2018. *Emotional AI: The Rise of Empathic Media.* SAGE, Thousand Oaks, CA.

[46] Mara Mills 2010. Deaf jam: From inscription to reproduction to information. *Social text* 28, 1 (Mar. 2010), 35-58. DOI: https://doi.org/10.1215/01642472-2009-059

[47] Mara Mills. 2015. Technology. In Rachel Adams, Benjamin Reiss, & David Serlin, eds. *Keywords for Disability Studies*, 176-180. New York University Press. New York, NY.

[48] Iain R. Murray and John L. Arnott. 1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America* 93, 2 (Feb. 1993), 1097-1108. DOI: https://doi.org/10.1121/1.405558

[49] MyHealthyApple Staff. 2022. Amazon Halo Tone Analysis : A Beginner's User Guide and review. *Myhealthyapple.com.* Retrieved December 5, 2022 from https://www.myhealthyapple.com/amazon-halo-tone-analysis-user-guide-review/

[50] Jeff Nagy. 2022. Autism and the making of emotion AI: Disability as resource for surveillance capitalism. *New media & society*, 0, 0 (Jul. 2022). DOI: https://doi.org/10.1177/14614448221109550

[51] Karen Nakamura. 2019. My algorithms have determined you're not human: AI-ML, reverse Turing-tests, and the disability experience. *UC Berkeley.* DOI: https://doi.org/10.1145/3308561.3353812

[52] The New York Times. 1925. Students Measure Fear by a Pupilometer, Kick Subject's Shins to Experiment on Anger. *The New York Times* (1925).

[53] Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B. Mariño. 2001. Speech emotion recognition using hidden Markov models. In *Eurospeech 2001: Seventh European Conference on Speech Communication and Technology*, Aalborg, Denmark. International Speech Communication Association, Baixas, France, 1-14. https://doi.org/10.21437/Eurospeech.2001

[54] Tin Lay Nwe, Say Wei Foo, and Liyange C. de Silva. 2003. Speech emotion recognition using hidden Markov models. *Speech Communication* 41, 4 (Nov. 2003), 603-623. DOI: https://doi.org/10.1016/S0167-6393(03)00099-2

[55] Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions.* Cambridge University Press, Cambridge, UK. DOI: https://doi.org/10.1017/CBO9780511571299

[56] Rosalind Picard (1995) *Affective Computing.* MIT Press, Cambridge, MA.

[57] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. arXiv:1810.02508. Retrieved from https://arxiv.org/abs/1810.02508

[58] Kat Roemmich, Tillie Rosenberg, Serena Fan, and Nazanin Andalibi. 2023. Values in emotion artificial intelligence hiring services: Technosolutions to organizational problems. In *Proceedings of the ACM in Human Computer Interaction (PACMHCI '23),* October 13 – 18, 2023, Minneapolis, MN, USA. ACM., New York, New York. Retrieved from https://www.nazaninandalibi.net/_files/ugd/63b293_88b25c6069b64b9bb8b67f1fc02dc281.pdf

[59] Sonja Rohrmann, Myriam N. Bechtoldt, Henrik Hopp, Volker Hodapp, and Dieter Zapf. 2011. Psychophysiological effects of emotional display rules and the moderating role of trait anger in a simulated call center. *Anxiety, stress & coping* 24, 4 (Nov. 2010), 421-438. DOI: https://doi.org/10.1080/10615806.2010.530262

[60] Sarah F. Rose. 2017. *No Right to Be Idle: The Invention of Disability, 1840s-1930s.* University of North Carolina Press, Chapel Hill, NC.

[61] James A. Russell. 1991. Culture and the categorization of emotions. *Psychological bulletin* 110, 3 (Nov. 1991), 426-450. DOI: https://doi.org/10.1037/0033-2909.110.3.426

[62] James A. Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review* 110, 1 (Jan. 2003), 145-172. DO: https://doi.org/10.1037/0033-

295X.110.1.145

[63] Luke Stark and Jesse Hoey. 2021. The ethics of emotion in artificial intelligence systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, March 3 – 10, 2021, Virtual Event, Canada. ACM., New York, New York ,782-793. https://doi.org/10.1145/3442188.3445939

[64] Klaus R. Scherer. 2009. The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion* 23 (Sep. 2009), 1307-1351. DOI: https://doi.org/10.1080/02699930902928969

[65] Björn W. Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM* 61, 5 (May 2018), 90-99. DOI: https://doi.org/10.1145/3129340

[66] Björn W. Schuller, Gerhard Rigoll, and Manfred Lang. 2003. Hidden Markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP '03)*, April 06 – 10, 2003, Hong Kong, China. IEEE., New York, New York, II-1-11-4. https://doi.org/10.1109/ICASSP.2003.1202279

[67] Björn W. Schuller, Gerhard Rigoll, and Manfred Lang. 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *2004 IEEE international conference on acoustics, speech, and signal processing (ICASSP '04)*, May 17 – 21, 2004, Montreal, Quebec, Canada. IEEE., New York, New York, I-577-I-580. https://doi.org/10.1109/ICASSP.2004.1326051

[68] Tobin Siebers. 2008. *Disability Theory*. University of Michigan Press, Ann Arbor, MI.

[69] Tom Simonite. 2018. This Call May Be Monitored for Tone and Emotion. *Wired*. Retrieved December 5, 2022 from https://www.wired.com/story/this-call-may-be-monitored-for-tone-and-emotion/

[70] Frank Thomas and Ollie Johnston. 1981. *Disney Animation: The Illusion of Life*. Abbeville Press, New York, NY.

[71] Silvan Tomkins. 1962. *Affect Imagery Consciousness: Volume I: The Positive Affects*. Springer Publishing Company, New York, NY.

[72] Silvan Tomkins. 1963. *Affect Imagery Consciousness: Volume I: The Negative Affects*. Springer Publishing Company, New York, NY.

[73] Voicesense. 2022. Mental Health and Wellness. *Voicesense*. Retrieved December 5, 2022 from https://www.voicesense.com/usecases/mental-health-and-wellness

[74] Meredith Whittaker, Meryl Alper, Cynthis L. Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, Marcel Salas, and Sarah Myers West. 2019. Disability, Bias, and AI. Retrieved December 5, 2022 from https://ainowinstitute.org/disabilitybiasai-2019.pdf

[75] William James. 1884. What is an emotion?. *Mind* 9, 34 (Apr. 1884), 188-205. Retrieved from https://www.jstor.org/stable/2246769

[76] Shoshana Zuboff. 2019. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs, New York, NY.