"I Didn't Know I Looked Angry": Characterizing Observed Emotion and Reported Affect at Work

Harmanpreet Kaur
harmank@umich.edu
University of Michigan
Ann Arbor, MI, USA

Daniel McDuff
damcduff@microsoft.com
Microsoft
Redmond, WA, USA

Alex C. Williams
acw@utk.edu
University of Tennessee
Knoxville, TN, USA

Jaime Teevan
teevan@microsoft.com
Microsoft
Redmond, WA, USA

Shamsi T. Iqbal
shamsi@microsoft.com
Microsoft
Redmond, WA, USA

ABSTRACT

With the growing prevalence of affective computing applications, Automatic Emotion Recognition (AER) technologies have garnered attention in both research and industry settings. Initially limited to speech-based applications, AER technologies now include analysis of facial landmarks to provide predicted probabilities of a common subset of emotions (e.g., anger, happiness) for faces observed in an image or video frame. In this paper, we study the relationship between AER outputs and self-reports of affect employed by prior work, in the context of information work at a technology company. We compare the continuous *observed emotion* output from an AER tool to discrete *reported affect* obtained via a one-day combined tool-use and diary study ($N = 15$). We provide empirical evidence showing that these signals do not completely align, and find that using additional workplace context only improves alignment up to 58.6%. These results suggest affect must be studied in the context it is being expressed, and observed emotion signal should not replace internal reported affect for affective computing applications.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *User studies*; *Walkthrough evaluations.*

KEYWORDS

Affect; emotion labeling; workplace

ACM Reference Format:
Harmanpreet Kaur, Daniel McDuff, Alex C. Williams, Jaime Teevan, and Shamsi T. Iqbal. 2022. "I Didn't Know I Looked Angry": Characterizing Observed Emotion and Reported Affect at Work. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3491102.3517453

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517453

1 INTRODUCTION

With affective computing applications on the rise, the use of Automatic Emotion Recognition (AER) tools—a form of AI technology used for observing people and their behavior in computational settings [81, 91, 123]—has become commonplace. AER tools have been applied in many contexts: to support learning via real-time monitoring of emotion, motivation and performance [137]; for diagnosis and treatment of emotionally-influenced diseases, and in tele-home healthcare systems [78, 126]; to support emotional development and socialization in children with autism [93]; to help people monitor or reflect on their emotions [77, 90, 140] and provide mood-based recommendations [130]; and to enable several other affect-oriented applications [118]. While AER technology initially used speech or text-based data, a facial analysis-based approach is thought to be a new frontier for obtaining continuous affect signals that can be applied in affective computing settings [12, 51]. However, before it can be used in real-world applications, it is important to understand how reliable AER technology is, to ensure that it is used responsibly and not assumed to be more accurate than it is.

Facial analysis-based AER technology uses facial shape and appearance in images or video frames to predict emotions expressed on faces [24, 121]. Recent work has critiqued this technology as following the common view of facial expressions, which assumes that "each emotion category is consistently and uniquely expressed with its own distinctive configuration of facial movements, which therefore can be used to diagnose its presence" [3]. This does not account for variability in emotion expression and perception, which are essential to the development and use of AER technology. As such, AER technology is applied without testing it in context. That is, without accounting for variation in either *internal* (e.g., history or current state, specific to an individual) or *external* (e.g., physical and temporal aspects of an individual's surroundings) context. Applying technology in this way is both negligent and unethical.

Our goal is to study the signals from AER technology within a specific external context—the workplace—to observe their ability to consistently label facial configurations. The workplace is a prominent application space for affective computing: information work is riddled with distractions that lead to stressful and unproductive work practices [58]; finding ways to keep information workers happy at work is key to unlocking productivity gains and overall well-being [47, 49, 67]. With the growing pressure remote work

is placing on people's wellbeing [57, 127], the need and opportunities for affective computing have become even more prominent. At the individual level, affective computing can support personal mood and goal tacking [89]. At the group level, AER used with people's consent and in aggregate formats can be applied as a feedback mechanism in, for example, remote presentations, to support more equitable turn-taking, and collaboration in team meetings, etc. [96, 111]. At the organization level, understanding individuals' job-related stress and factors causing this stress can provide a necessary picture of how organization level stressors can be measured, and subsequently addressed [135]. Typically, in workplace settings, self-reports are the most commonly used approach for collecting subjective perceptions of affect. However, self-reports place a high burden on people and become infeasible to collect in most real-world long-term applications (outside of a research study), because people simply do not have the time to report how they feel on a regular basis. Moreover, self-reports can be riddled with biases of their own: relying entirely on them assumes that people can consistently and accurately report their internal feelings.

In this paper, we use the signal from an existing, state-of-the-art AER tool (*observed emotion*) and the signal from self-reports (*reported affect*) to provide a comparison between these two current ways of perceiving affective states. In prior work, affective states are usually tracked via self-reports that use, for example, the PANAS scale [134] to establish people's feelings (e.g., [83]). This is both cumbersome and not a continuous signal, thus paving the way for applying AER tools as an alternative. To evaluate the use of AER technology in the workplace, we present results from a day-long combined tool-use and diary study ($N = 15$, total 104 hours of tool data, 331 diary entries). We observed people's emotion and workplace context for a day using an existing AER and context logging tool. Participants maintained a diary for the same day to provide subjective labels of affect and productivity, in 30-minute intervals throughout the day. We also conducted follow-up interviews with the participants to help us understand their internal context and qualify the nuanced differences between the two data sources. In light of no existing objective ground truth mechanism, a comparative approach such as ours is common for establishing convergent or discriminant validity among different signals (e.g., prior work in predicting mental health states using data from online communities takes a similar approach [9, 32]).

Our results show that outputs from the AER and context logging tool that represent the dominant emotion for a timeframe cannot successfully capture people's emotions in the workplace. The external workplace context is an important differentiator between the training datasets used for AER technology (which are comprised of posed expressions captured in in-person settings) and people's lack of outward emotional expression in front of a computer screen in the workplace. We show that people have unique profiles for dominant observed emotions—a common metric used in prior work that predicts an emotion label for a timeframe based on the emotion with the highest magnitude throughout that timeframe. Similar unique profiles exist when we calculate the valence of reported affect, but these two types of affect profiles do not align well (35.4% match). We introduce two new metrics to reflect a different type of observed emotion profile: *emotion spikes*—emotion labels that differ from their expected mean—and *baseline emotions*—the most prevalent emotion labels throughout a given timeframe. We find that an observed emotion profile based on these two metrics is better aligned with reported affect in the workplace (58.6% match), but only after including several contextual factors about the participants' task profiles and workplace activity. We discuss the implications of misalignment between these two affect signals for future applications of AER tools. Particularly for the workplace context, we discuss whether alignment between AER and people's true internal affective states can—or even should—be a goal for research in this space. We also highlight high-level themes from our experience in attempting construct validity in this complex facial expression domain. Overall, this paper contributes:

- A comprehensive study that characterizes the use of AER technology in the workplace context,
- Two new metrics—emotion spikes and baseline emotions—that better represent observed emotion in the workplace context compared to the commonly used metric that relies on dominant emotions.
- A characterization of the magnitude and type of misalignment between observed emotion and reported affect, two prominent signals for ascertaining people's affective states.
- A discussion of construct validity in the facial expression domain.

2 RELATED WORK

2.1 Facial Expressions and Emotion

As early as two millennia ago scientists and philosophers proposed the notion that faces communicated consistent information about a person's state [33]. Darwin's book on "The expression of the emotions in man and animals" [19] further established this idea. By the 1960s, psychologists began rigorously validating these hypotheses and found empirical evidence that people across cultures interpreted facial muscle configurations similarly for a set of "basic" emotions [30, 59]. These seven emotions—anger, contempt, disgust, fear, happiness, sadness, and surprise—were termed "basic" because: (1) they had distinctive states such that they could be differentiated from one another, and (2) their biological contribution included shaping both unique and common physical features that these emotions displayed, and their functional role in evolution [28]. Tracy and Randles provide an overview of the four models of basic emotions, noting similarities and discrepancies in which emotions are considered basic in various models [131]. Ekman's model is the most commonly used [27], especially given the use of the seven emotions mentioned above for mapping to the corresponding facial movements in the Facial Action Coding System (FACS) [29, 38].

These studies conducted by psychologists have been used as evidence of the universality thesis by some researchers [88, 103] and hotly contested by others [3, 34, 108]. Over the past 30 years, whether it has been proven that expressions of emotions are indeed universal has been a source of considerable debate [3, 34, 40, 60, 108], with some arguing that these cross-cultural studies had methodological flaws. As Kappas states in his critique of this presumed universality, "any demonstration of the non-exclusiveness of this relationship [between facial expressions and universal emotions] is sufficient to cast doubt on the immediate diagnostic value of facial

actions as indicators of affective state" [64]. With several studies calling this universality into question, other models of emotion have garnered some attention: (1) the appraisal view, wherein affect is modeled after individual goals and needs, and assigned subjective metrics and values; and (2) the dimensional view, where affect is treated as a consequence of the neurophysiological system [20, 107], most commonly combined dimensions being that of valence (positive or negative) and activation or arousal [8]. While these models are helpful, adopting them for applications is challenging due to their high-level treatment of affect [64]. As such, Ekman and others have backed the basic view of emotions—particularly in the context of understanding emotions for applications (e.g., affective computing)—with a simple observation: universality is not a necessary condition; it is undeniable that facial expressions contain rich information about a person's affective state.

2.1.1 Deliberate vs. Spontaneous Emotion Expression. One common differentiation made in determining emotion from facial expressions is grounded in whether the expression was *deliberate* (i.e., posed or made with a certain circumstance in mind, such as interpersonal regulation) or *spontaneous* (i.e., unmodulated and elicited without conscious thought) [26]. This question is has been prevalent for expressions of positive emotions in particular: Frank et al. [37], Ekman [28], and several others (e.g., [54]) acknowledge this difference as the Duchenne smile (spontaneous) which shows true enjoyment compared to a fabricated, social (deliberate) smile. Research (e.g., [72, 125]) has shown that these smiles can also be distinguished by people. More recently, Namba et al. found that other emotions such as amusement, disgust, sadness, and surprise, can also be differentiated by people into the "shown" and "felt" expressions of these emotions [97].

As emotion recognition technology becomes a reality, deliberate vs. spontaneous emotion expression is critical in two ways. First, there is the consideration of whether technology can distinguish between these two types of expressions. There have been some successful attempts with training automated approaches using temporal factors (e.g., onset and offset duration [13]), lip corner movement speed and magnitude [117], brow actions (e.g., their velocity, duration, and order of occurrence [132]), a combination of these [22], and even including contextual factors based on optical flow over the face for a period of time [55]. Perusquía-Hernández provides an overview of automated techniques that rely on human annotation, bio-signals, and sensors [105]. Second, there is the question of how deliberate and spontaneous expressions could lead to varying outputs when used for training emotion recognition technology [5, 87]. Studies conducted by Cordaro et al. [14], Elfenbein [31], Durán et al. [25], and others, and their comprehensive review by Barrett et al. [3] provide empirical evidence that deliberate vs. spontaneous emotion expressions lead to different outcomes. There is consensus that this affects AER, but characterizing exactly how that happens and how it changes with varying contexts (e.g., face-to-face, workplace, etc.) is an important open question.

2.1.2 Gender Differences in Emotion Expression. Dating back to the theories of construction of family roles, men and women were considered to have different facial expressions and emoting capabilities. Parson et al.'s famous theory on family roles [104] cites women as the "expressive experts," and men as the "instrumental experts," noting women's supposed ability to emote and empathize better than men [74, 120]. Women tend to perform better than men in matching people's facial expressions to how they are feeling [52, 128]. Studies have also shown differences in how men and women emote, captured via facial expressions, in response to the same stimuli, even when they perceived the stimuli similarly [23, 52]. In general, self-report and observational studies have found that women smile more and express more positive valence emotions, whereas men show more anger and aggression [7, 35]. We compare to these gender differences using data from our participants, but note a critique of this prior research that it does not include emotion profiles for the broader spectrum of gender identities [68]. Our participant pool is similarly divided into men and women (self-reported, open-text response); we note this as a limitation.

2.2 Automated Emotion Recognition

Advances in facial detection and analysis have enabled technologies that leverage computer vision to detect prototypical expressions of the basic emotions (joy, anger, fear, sadness, surprise, disgust, contempt) from images [4, 92], and these give us useful signals, especially when tracked longitudinally. Recent multimodal affect sensing technologies rely on video and audio signals instead of wearable sensors. Tools that use computer vision typically apply them for facial detection and 3D head position estimation, landmark extraction, and facial expression analysis and modeling, and output predicted probabilities for different emotions based on this data. Examples include MultiSense, which analyzes people's affect and non-verbal cues (via audio and video signals) for assisting in mental health settings [123], and Mansoorizadeh and Charkari's approach that uses multimodal information fusion of several facial signals to recognize emotions [81]. D'Mello and Kory [24] and, more recently, Soleymani et al. [121] provide a comprehensive survey of existing tools and technologies for affect sensing.

2.2.1 Critiques of Existing Automated Approaches. Durán et al. provide an extensive meta-review of the psychology research on coherence between emotions and facial expressions (see also Section 2.1 above), to test "the implicit assumption that facial expressions co-occur with emotions" [25]. Their review shows that there is low coherence between the two. Building on that, Barrett et al. critique AER technology because of its reliance on a common view of facial expression [3]. Barrett et al. provide a comprehensive review of existing work on emotion expression and perception—evidence from cross-cultural studies of healthy adults, newborns and young children, and people who are congenitally blind—to conclude that the common view of facial expression does not account for the contextual and variable nature of facial expressions. Thus, both Durán et al. and Barrett et al. provide evidence for the problematic reverse inference applied in AER technology: while facial configurations can be captured reasonably well, there is no reliable way of predicting emotions from these facial configurations.

Barrett et al. propose four metrics for evaluating AER technology that must be met before this technology can be used widely: reliability, specificity, generalizability, and validity. Using evidence from cross-cultural studies of healthy adults, infants, and children, and

studies of congenitally blind individuals, they show that AER technology currently has limited reliability (the same emotion cannot be reliably expressed or perceived from a set of facial landmarks), a lack of specificity (no one-to-one mapping between facial configurations and emotion categories), limited generalizability (there have been no contextual or cross-cultural studies evaluating this technology), and as a result, limited validity [3]. Samadiani et al. provide guidance for why AER technology fails on these metrics: variation in illumination levels which affects the accuracy of extracting facial features, head post, and subject dependence [110].

In our work, we seek to study AER technology in a contextual setting, thus following the research guidelines from Barrett et al.'s review [3]. By studying AER technology in a constrained external context—workplace—and capturing nuanced internal context—by asking participants to maintain diaries and conducting follow-up interviews—we hope to provide a comprehensive evaluation of AER technology in a particular setting.

2.3 Understanding User Affect In Situ

2.3.1 Affect Signals in the Workplace. Affect recognition has come a long way from categorization into basic emotions based on posed data acquired in lab settings [1] to understanding affect in situ using a combination of subtle context-specific factors (e.g., head shakes and shoulder movements) along with recorded facial expressions [98]. Müller and Fritz establish another important relationship that is associated with affect, that of emotion with task progress, for software developers [95]. This understanding of affect and how it changes based on signals from various sensor data has enabled multiple productivity-centric applications around affective states. For example, AffectAura [90] uses a multimodal desktop-based and body-worn sensor setup to predict the valence, arousal, and engagement components of affect state for people, so that they can reflect on their days from an affective perspective. Similarly, MoodWings [80] captures affect states from sensors when people are stressed, and uses physical interfaces (wings) to warn them of their stress levels and helps them moderate. Kapoor and Picard [63] use a multi-sensor affect recognition system that relies on facial expressions and postural shifts to understand and improve computer-based learning environments for students.

2.3.2 Benefits of Understanding Affect in the Workplace. The consideration of well-being in the workplace has received special attention in the literature on Organizational Science, leading to research areas like positive organizational scholarship [11] and behavior [79]. Prior work on organizational behavior has found that a happy disposition at work leads to more productive outcomes, measured via performance ratings [49, 136] or through cognitive judgements captured via ratings [36, 73]. This "happy-productive worker" hypothesis was corroborated for information workers by [21, 47, 82, 102], showing that despite fragmented work, software developers can solve analytical problems and resolve issues tracked on open repositories in shorter timeframes when they feel positive about work. To that end, applications designed for well-being in the workplace often rely on self-reported affect values via the Positive and Negative Affect Scale (e.g., [83, 85, 86]) and time spent on task or number of errors for task performance [2].

Affective data has been incorporated in to provide feedback to meeting attendees [112, 113] and in presentation software to highlight audience reactions [96]. These studies have shown that carefully designed affective feedback, measured using automated tools, can help improve remote communication and collaboration. With the opportunity for continuous affect signal provided by AER technology, workplace applications that rely on this technology are on the rise. Before that can happen at a large scale, we seek to evaluate this technology in a contextual way. We thus examine data from AER technology embedded in a multimodal emotion and context logging tool [92], along with affect values from people in a diary study, to compare the two sources and test the feasibility of using AER tools in workplace affect and productivity applications.

3 STUDY DESIGN

We conducted a combined tool-use and diary study to understand how affect manifests during a regular workday as information workers go through their daily tasks on a computer. Our goal was to compare these two sources of affect data to investigate whether AER produces measurements that are consistent with the participants subjective experience and to test the feasibility of using facial expression prediction for continually monitoring affect in the workplace. We compare two types of data around how people feel:

Observed Emotions: External emotion labels predicted by an Automatic Emotion Recognition (AER) tool [92] based on the facial landmarks observed for a person; recorded every microsecond.

Reported Affect: Internal, self-reported overall feeling, recorded by participants in their diaries, every 30 minutes.

3.1 Study Participants

The combined tool-use and diary study enables an understanding of how the observed emotion data compares to the reported affect data, and how affect manifests around different types of workplace activities. We recruited 15 participants (9 women, 6 men; self-reported gender profiles) at a large technology company to keep a diary about their workday for one day of the week when they had 0-2 meetings. Participants had mostly research-focused roles, ranging from researchers and developers to research interns, whose workday largely comprised of activities carried out on a computer. Participants installed an existing, state-of-the-art AER and context logging tool developed by McDuff et al. [92] on their workstation desktops[1]. They were provided a Microsoft LifeCam HD-300 (resolution: 1280 x 720 px) for the study. These webcams were placed on top of their computer monitor, pointing down at an angle of ~30 degrees, and connected to their computer via USB 2.0. All participants had offices in the same building and thus lighting was similar, a combination of overhead florescent lights and natural light from windows. We ensured that no participant had poor illumination or was back-lit during the study. The cameras were not otherwise calibrated or adjusted.

[1]The AER and context logging tool was only installed on peoples' desktops. Although some people also used laptops (e.g., during meetings away from their workstations), these devices did not have the necessary computational power for the tool to operate without making other applications lag. This computational power was required by the tool to output observed emotion and context logs in real time, without storing any audio or video data, thus making it privacy-preserving compared to other alternatives.

The tool collected emotion and context data in the background, while participants maintained a diary throughout the workday. All participants were compensated with $20 lunch coupons for participating in our study. The first author conducted an in-person interview a day later, showing the participants their affect and activity patterns with the goal of validation and clarification of observed data and interpretations of observed behaviors, in comparison to their reported affect. We opted to collect data for a single day to reduce the load on the participants to provide diary entries and to ensure that we were getting high quality data. We also wanted to maximize people's recall of their internal context for the retrospective process of the follow-up interviews—a longer diary study would have made this challenging. Despite our reasoning for this setup, we acknowledge the short timeframe of our diary study as a limitation. The study was reviewed and approved by the IRB.

3.2 AER Tool for Capturing Observed Emotion and Context

Observed emotions were collected using an existing AER tool developed by McDuff et al. [92][2]. This tool has been deployed in workplace settings in the past [67], and has been successfully used in several studies that rely on AER and context data [18, 70, 89]. It captures facial expressions and workplace context in microsecond intervals and processes the data locally to provide emotion labels and context logs. We selected this particular tool for our study for three reasons: (1) it relies on state-of-the-art AER technology to provide observed emotion labels (see visual pipeline below); (2) it records workplace context data (e.g., mouse and keyboard usage rates, tabs and window switches), which allowed us to study the use of AER technology in context without additional development costs (see context pipeline below); and (3) it is privacy-preserving in that the observed emotion labels are computed locally in real time—no audio or video is recorded. Below, we describe the relevant pipelines of this tool. For more details on the design decisions behind the various features of the tool, please refer to [92].

3.2.1 Visual Pipeline. The visual pipeline utilizes a webcam and processes the video frames in real time. Since the signal processing is done in real time, the tool never stores full videos or image frames, providing the user with a greater degree of privacy. It only stores the output values from the models that run on this data. It detects faces within the video feed using a convolutional neural network (CNN) detector and extracts landmark positions of key facial features. Multiple faces (up to 5) can be detected at a time. The tool is able to capture these signals even when someone is also using the camera or microphone in another application (e.g., a video call). The distance of the user's face from the camera is extracted using the inter-ocular distance calculated from the facial landmarks, and facial regions of interest are analyzed using an emotion detection algorithm.

The emotion detector returns eight probabilities for each of the following basic emotional expressions: anger, disgust, fear, joy, sadness, surprise, contempt and neutral, with an accuracy of ~87%. The tool assigns a 0–1 probability value to each non-neutral emotion and "neutral" is calculated as the (1 - ∑(all-other-emotion-probabilities)) for each microsecond. This emotion detector follows the AER algorithm developed by Barsoum et al. [4]. It is a deep convolutional neural network (DCNN) architecture which uses probabilistic label drawing to output the eight emotion probabilities, based on facial landmarks and movement captured by the tool. Ideally, AER algorithms would be trained on ground truth emotion labels or FACS (Facial Action Coding System) codes obtained based on facial analysis performed by trained humans. Given the large amounts of data to consider and the challenges of training individuals for this task, crowd-sourced labels have been proposed as an alternative. These crowd-sourced labels can be understandably noisy, given what we know about the psychology of recognizing emotions from facial expressions (see Section 2). Barsoum et al.'s probabilistic approach provides multiple labels (as relevant) for each facial expression with the corresponding probability values. For example, an expression can be labeled as 75% fear and 25% anger instead of giving it one label. Given it's high accuracy (~87%), this AER algorithm and DCNN architecture has been employed in several state-of-the-art AER tools (e.g., [16, 69, 76]), including the one we used for our study [92]. The detector used by the tool is publicly available (EmotionAPI[3]), allowing other researchers to replicate this method. For more information on the performance of the facial expression classification, see [4].

3.2.2 Context Pipeline. The tool tracks information about open applications and interactions with computer peripherals. It records when an application is opened, closed, in focus (the front application), minimized, or maximized, with the corresponding timestamp. It also logs mouse movements and clicks, and keyboard inputs. These actions provide a rich log of the participants activities.

3.2.3 Observed Emotion and Context Data. The tool collected a total of 104 hours of observed emotion and context data, with 7 hours per person, on average (min=4 hrs, max=9.7 hrs). We analyzed the continuous data by plotting the AER and context data aggregated from microseconds to over 5-minute intervals for the entire day. 5-minute intervals were selected after a grid search ranging 1-10, 15, 20, 25, and 30 minute intervals, optimized for consistency in the recorded emotions—the selected interval represents the average duration for which emotion profiles did not change. The aggregation function calculates the mean over every 5 minutes of data after removing outliers (values that are two standard deviations away from the mean). These outliers were rare and only present in the context data in cases such as when a participant accidentally long-pressed their keyboard keys because they were writing in a notebook that was lying on top of the keyboard, or when a participant left their desktop unlocked while at lunch with music playing on YouTube in the background. We were able to determine that these were outliers based on corresponding diary entries from the time periods in question and confirmed these instances during the follow-up interviews. We also excluded any data recorded for additional faces in the environment, since shared emotion observation was out of scope for our study. This happened for 4 out of 15 participants since they had a shared or open office space. After cleaning up the data, we used these 5-minute aggregate observed emotion and context values to find patterns across participants.

[2]We requested and obtained access to this tool for our study.

[3]Microsoft, Inc.

Current Time	Task Start Time	Task End Time	Task Description	Task Urgency [1-low -- 7-high]	Current Windows being used for the task	Other Windows being used	Any breaks away from the computer? (e.g., restroom, kitchen, meetings)	Any breaks on the computer or phone? (e.g., social media, youtube, reddit)	Other Interruptions? (e.g., email, people stopping by your office)	Overall Feeling
8:45	8:45	9:05	Settling into the office	1	Outlook					
9:00			Reading about affinity diagramming	5	Chrome	Edge	Coffee Break			Irritated with the loud noises outside the office
9:30		9:40	Reading about affinity diagramming	5	Chrome	Edge				Motivated
9:40			Working on bot, reading documentation	6	Chrome	Outlook	Outlook, looking at the new team pictures		Checked Outlook	Intimidated
10:00			Prototyping for bot, reading documentation	6	Chrome, Teams	Outlook	Bathroom break	Texting friends about weekend plans, texting friend about lunch plans		In the zone

Table 1: The first five entries of a diary, shared with participant consent.

3.3 Diaries for Capturing Reported Affect and Context

The main purpose of the diaries was to record similar data as the AER tool, but directly from people as subjective feedback. Throughout the study day, participants recorded their overall feeling for a given time interval along with workplace context such as task description, urgency, the windows and applications being used, any physical or digital breaks during that interval, other interruptions, etc. Table 1 shows part of a diary from one of our participants, shared with their consent. Participants were provided a diary template and instructed to add diary entries every 30 minutes and any time they started a new task. Given that 30 minutes is the norm for calendar-based scheduling and task planning in corporate workplace contexts, we believed that it would be least disruptive to people's work routine to maintain a diary in those intervals. We encouraged people to think of tasks as fine-grained and note a task switch when they started using new applications (e.g., web browser vs. programming IDEs) or used the same applications for different reasons (e.g., using a web browser for literature search vs. collaborative writing, switching code branches for programming different elements). Without developing a tool to support these diary reports, there was no way for us to ensure that participants recorded all possible task switches or made diary entries every 30 minutes. Although most participants followed our instructions, we note this as a limitation of our chosen methodology. Participants also rated their affective state using the Positive and Negative Affect Scale (PANAS) [134] at the beginning and the end of the day. PANAS has been extensively used in prior work to measure affect in the workplace (e.g., in [83–85]) and other contexts (e.g., online support groups [50], behavior change chatbots [41], driver-vehicle interfaces [39]). Further, PANAS has been shown to be a reliable measure of positive and negative affect, with people using it to report affect in a way that is internally consistent with their mood [134].

3.3.1 Reported Affect and Context Data. We recorded 331 diary entries in total, 22 per participant, on average (min=16, max=40). To obtain each participant's reported affect, we classify the *valence* of the reported feeling. Valence represents whether a feeling is positive, negative, or neutral, and helps us classify the various terms used by our participants to express emotions into a discrete set of categories. We code reported feelings as positive or negative valence when they have clear indications such as feeling "good", "enthusiastic", "inspired"; or "disappointed", "irritated", "not good." We also take into consideration the task and context information from participants' diaries when the valence of their reported feeling is not outwardly clear. The first and last author coded all entries with substantial agreement (inter-rater reliability of 0.77 measured using Cohen's Kappa); any differences were clarified via discussion.

For all participants, a majority of the PANAS items (out of 20) were neutral (i.e., no change) between the beginning and the end of the day. We used the non-neutral items from the PANAS data to validate the overall valence of reported affect by checking that the valence of the non-neutral PANAS items (positive or negative) matched the valence from the diaries. In this way, we not only validated that the diary entries matched the PANAS signal, but also noted why the more granular snapshot from the diaries was needed for reported affect: PANAS scale values alone would have presented an overall neutral affective state than the granular reports from the diary because the positive and negative valences counter-balanced each other at the day-level.

3.4 Follow-up Interviews

We conducted artifact-based interviews with the participants a day after they completed the combined tool-use and diary study. Prior to the interview, we analyzed and created visualizations of the data collected from the AER tool, and annotated the timestamps with the diary entries. We superimposed people's diary entries on output visualizations about emotion labels, mouse and keyboard usage, window and tab switches, and people's distance from the screen [4]. Participants were then asked to do a retrospective walk-through of their previous day and evaluate these visualizations based on their recollections, as a whole. We used printouts of visualizations for better accessibility for people.

During the interview, we focused on questions pertaining to (1) how well the observed log data matched the reported diary data; (2) asking for explanations where the log data and diary data were conflicted; (3) asking for potential applications of AER tools and how comfortable people felt with such a tool running in the background;

[4] An anonymized example of these visualizations is included as supplementary material

and (4) general questions about work habits, focus, and stress, and how these manifested for the individual. Our interviews were ~45 minutes long, on average. While the interview protocol was semi-structured, the authors prepared the same set of visualizations for all interviews. All open-ended data generated from the interviews was analyzed using inductive thematic analysis [6], by first applying open codes followed by axial coding using affinity diagramming.

3.5 Temporal Matching of Observed Emotion and Reported Affect for Comparison

There were two uses for temporal matching of the observed emotion and reported affect signals: (1) for the visualizations created for the follow-up interviews, and (2) for the direct statistical comparisons between the two signals. For the former, we used 5-minute intervals for plotting the AER tool data in the form of bar and line charts. The diary entries were simply superimposed for the entire relevant duration using brackets and text (see the anonymized example included as supplementary material).

For the direct statistical comparisons between the AER tool (observed emotion) and diary (reported affect) data, we updated both signals to the same level of granularity using valence. The reported affect data was already coded using valence (Section 3.3.1). For the observed emotion data, we encoded anger, contempt, disgust, fear, and sadness with negative valence; and happiness and surprise with positive valence. This valence classification is pretty standard, except for "surprise." Prior work has shown that surprise can have a positive or negative valence [99], depending on how a person responds to the surprising event [27]. For the workplace context, there has been some indication that surprise can be positive given its link to satisfaction [122, 133]. Our follow-up interviews corroborated this positive valence for surprise for our context. With these changes, each 5-minute interval of observed emotion was represented as positive, negative, or neutral valence. Note that we did not have any control over the granularity of the original AER outputs—most AER technology, including the one employed by the tool we used for our study, follows a basic view of emotions [27] and provides the seven basic emotions as output labels.

Once at the same level of granularity, we compared the reported affect from diary entries for a time interval to the closest equivalent 5-minute interval from the AER tool. In this way, we avoided making assumptions about how someone reported affect beyond the particular time instance for which it was reported. For example, it is not necessary that someone who reported feeling happy about their task at 9:30am felt happy for the entirety of the 9am to 9:30am interval. As such, we compared the 9:30am diary entry to the observed emotion aggregated over the 9:25-9:30am interval. We used chi-squared tests for these comparisons due to the discrete nature of the diary entries, as noted above. We also report effect sizes using Cramer's V, which is commonly used for contingency tables of sizes greater than 2x2, like ours.

Of the 331 possible comparison points, there were 215 entries for which both types of affect data were available. Although we had asked participants to record their task progress and affective state every 30 mins, there were some instances when they forgot to do so. There were other times when the participants were working away from their desktops (e.g., using their laptops at a coffee shop or in meetings), and the AER tool was only available on their workstation desktops. Due to these inconsistencies, there were instances for which we only had one affect signal. All direct comparisons are made using chi-squared tests based on the 215 entries for which both signals were available.

4 RESULTS

Our goals for the study were three-fold. First, we wanted to evaluate the use of AER tools within specific internal and external contexts, per Barrett et al.'s recommendation [3]. The diary and the follow-up interviews shed some light on people's internal context, and the workplace served as the external context. Second, we wanted to characterize the two ways of capturing affective states in the workplace: one from the AER tool and the other from self-reports. We thus include an in-depth analysis of these signals and our process for identifying patterns in this data. Finally, we wanted to compare the two types of affect signals to establish convergent or discriminant validity. Our results below are structured with these goals in mind: qualitative evaluation of the tool by participants, characterization of observed emotion states, characterization of reported affect states, a comparison between the two signals, and our attempts to further establish convergent or discriminant validity by using workplace context to shape affect signals.

4.1 Qualitative Evaluation of the AER Tool

The day after the combined tool-use and diary study, we conducted an artifact-based retrospective, semi-structured interview with participants to go over their previous day (when the data was collected). Of the 215 datapoints for which we had information from both the diary and the tool, participants subjectively evaluated 183 (85%) datapoints as being accurately characterized by the AER tool. Note that this evaluation is based on both the emotion label assigned to people's facial movements by the tool, as well as the workplace context recorded by it (e.g., mouse and keyboard usage, leaning in towards or backing away from the screen).

According to the participants, the tool performed better than they expected in gauging their facial movements in the workplace. Most participants had very low expectations going into the experiment: "I'm expecting a lot of noise in this data" (P4) and "I'm curious to see the results...I have no idea if this will work" (P11). After seeing the resulting visualizations from the tool superimposed with their diary entries, participants were surprised by how well the tool was able to perceive and label their facial movements in front of a computer screen. However, we also consider this low expectation prior to seeing the tool's outputs as a caveat for contextualizing the 85% accuracy number from subjective evaluations noted above. That is, we cannot be sure how any bias from the tool surpassing people's expectations affected these evaluations. Below, we highlight some pros and cons of the tool as mentioned by the participants.

4.1.1 High Perceived Reliability. Participants found that the tool captured a nuanced signal, using a combination of the emotion labels for facial movement and workplace context: "I'm surprised by how much my diary entries match the output emotion label and the activity monitoring" (P7). They noted that this was an important and accurate signal, even when the tool did not have the right label for it. As a result, they brought up opportunities for how they

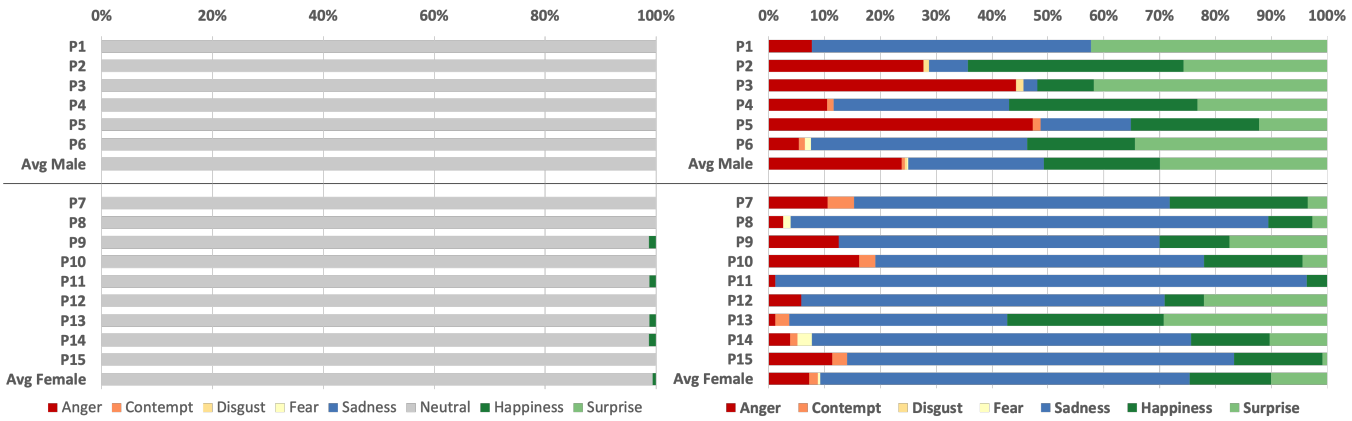

Figure 1: Dominant emotions per participant output by the AER tool. An emotion is "dominant" if it is the top predicted emotion for the facial landmarks for a given timeframe [66]. Left: Dominant emotion profiles based on all eight emotions shows "neutral" as the dominant emotion overwhelmingly across all participants. Using this dominant emotion output without further processing does not capture the subtle emotion signals that could be extracted from the non-neutral emotions band. In contexts where people do not naturally emote (e.g., at work, in front of a screen), capturing these subtle signals could be critical for effective use of AER technology. Right: Dominant emotion profiles after excluding "neutral" class label.

might control the internal labeling for different facial movements: "At least it consistently catches me yawning in front of my screen as 'surprise'... I wonder if you could change that encoding" (P15).

4.1.2 Low Specificity for Some Emotions. Given that the tool relied on only eight emotion labels to characterize a facial configuration—a common critique of AER tools [3])—there were times when it could not distinguish between configurations that are similar across a wide range of emotions. The tool did not have the right level of specificity to describe these negative results. P15 explained high levels of surprise in their data as being representative of when they were tired: the tool recorded their tired state as "surprise," because the yawning expression is pretty similar to one of exaggerated surprise (mouth wide open and eyebrows raised). Similarly, P5 noted that their data showed a lot of "anger" emotion label. They explained this as:

> "I am short, and I sit further away from my screen than most people, because I fidget too much. So I am always kinda looking up at my screen... like my eyeballs are towards the top end of my eyes. I can see why that expression looks angry. In fact, even my friends tell me I look angry when I'm working. But, obviously, I'm not *always* angry. [I'm just working.]"

4.1.3 Missing Context for Idiosyncratic Behaviors. While most people agreed that the observed emotions detected by the AER tool were a mostly-accurate representation of their facial expressions as they worked, they wished that the tool could also capture the subtleties of their work routine and adjust accordingly. For example, when asked about the relatively high number of spikes in observed contempt in their data, P7 noted that:

> "I fidget a lot. Also every time someone walks by my office, my glance automatically goes in that direction. I think the logger was too sensitive to that, and maybe that sideways, straight-faced expression is showing up as contempt."

4.1.4 Opportunities for Self-Tracking. Most participants noted that seeing the output from the tool was a helpful gauge for their day, and they would like to see it built into a personal tracking tool, one that they controlled: "It's weird how much these visualizations make sense. Like, I'm weirded out by it in a good way. I want to see more of this so I can track my own mood and goals..." (P2). They mentioned that the tool could serve as "daily work journals or capturing daily thoughts about my work" (P7). Of course, these participants also noted that they would want control over their data and not want to share it with any organization.

4.1.5 Privacy and Autonomy. When asked about the potential applications of these AER tools in the workplace, participants were open to trying the tool, as long as they had absolute control over it: "Honestly, I would be so worried about surveillance that I would not download this tool unless I could ensure that the data could be accessed by me alone. I like seeing these visualizations, but I don't like the opportunity for surveillance if a tool like this one was mandated by organizations" (P6). Participants also wanted autonomy over the classification labels assigned to their facial movements. P15 noted that "if I could just fix the label for every instance—whenever I'm showing surprise it's actually yawning—this works. It's capturing the right signal, but mislabeling it."

4.2 Understanding Observed Emotion Labels

4.2.1 Dominant Emotions. In our analysis of the AER data, we find that while the detected emotions change quite frequently, there is usually a single emotion that dominates any given time frame, which we identify as the *dominant emotion*. Prior work defines "dominant emotion" as the most prominent emotion seen at a given timeframe [15, 66]. For our 5-minute aggregation intervals for AER labels, we identify the emotion label with the highest magnitude, on average, over the 5-minute period as the dominant emotion.

Participant	Anger	Contempt	Disgust	Fear	Sadness	Neutral	Happiness	Surprise	Baseline Emotion
P1	8	13.5	6	**25**	**44**	100	11.5	**31**	Neutral, Sadness, Surprise
P2	4	11	3	4	13	100	**23**	**19**	Neutral
P3	7.6	7.6	7.6	14	7.6	100	**20**	**16.5**	Neutral
P4	5	6	2	3.5	**25.6**	100	**30**	**23**	Neutral
P5	**36.5**	4	11	7	12	100	**38**	13.5	Anger, Neutral
P6	3	1	4	4	**19.4**	100	**22.6**	**21.5**	Neutral
Avg Male	10.6	7	5.6	9.6	**21**	100	**24**	**21**	
P7	7	12	6	7	**26**	100	**27**	13	Neutral, Sadness
P8	6.6	17	5	8	**56.6**	100	**37**	**22.4**	Neutral, Sadness
P9	7.5	15	10	10	**37.5**	100	**17.5**	**20**	Neutral, Sadness, Surprise
P10	4.4	3	**16**	3	**48.5**	100	**31**	**16**	Neutral, Sadness
P11	**43**	7	6	7	**85.4**	100	**43**	7	Happiness, Neutral, Sadness
P12	3.5	13	2	8	**59**	100	**23**	**29**	Neutral, Sadness, Surprise
P13	11	8.5	6	6	**61**	100	**57**	**50**	Happiness, Neutral, Sadness, Surprise
P14	14	**15.4**	5	23	**69**	100	**37**	**23**	Neutral, Sadness
P15	23	12	2	3.5	**53.5**	100	**42**	11.4	Neutral, Sadness
Avg Female	13	11.5	6.5	8	**55**	100	**35**	**21**	

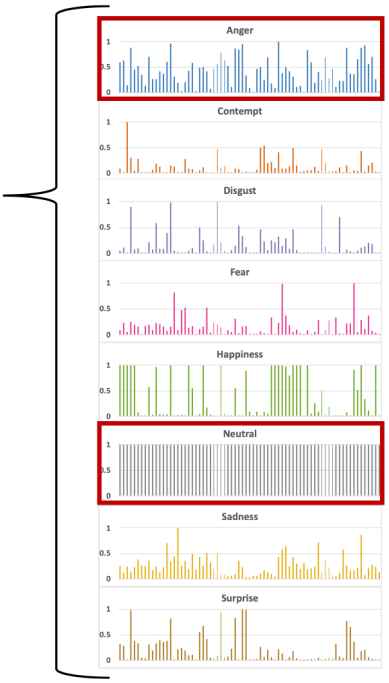

Figure 2: Left: The percentage of times each emotion spiked per participant (each emotion's value is out of 100%). Right: Normalized values per emotion for Participant 5. Baseline emotions picked as those that are prevalent throughout the day.

We calculate the number of times each emotion is dominant throughout the day for each participant, which gives us an observed emotion profile (see Figure 1). Figure 1-Left reveals that for all of our participants, "neutral" is the emotion that is predicted based on their facial landmarks throughout the day. In fact, for a majority of the participants (11 out of 15), "neutral" is the dominant emotion throughout the day, for all time intervals. Recall the presence of "neutral" is effectively the absence of other non-neutral emotion outputs given that it is calculated as such. We hypothesize that this dominance of "neutral" is in part due to the training data used for AER which relies on in-person, posed facial expressions ascribed to each emotion by a selected group of people. As noted in prior critiques of AER technology, the change in context from in-person to workplace can affect how people emote [3, 25].

Indeed, our workplace context is a setting where people do not naturally emote, i.e., in front of their computer screens. Our goal is to to understand how sensitive AER is in capturing emotions in this context. To enable this, we look for changes in the dominant emotion profiles after excluding the emotion label "neutral." The AER outputs are obtained at the microsecond level and then aggregated by us at 5-minute intervals. Imagine a 5-minute interval during which neutral is primarily the emotion recorded, with a couple of microseconds of anger, followed by surprise and happiness. If we were to simply aggregate the AER output based on the dominant emotion, we would register this entire interval as "neutral". By excluding the overwhelming presence of the "neutral" emotion label, we are able to observe dominant emotion profiles that highlight the subtle micro-expressions of emotions we might have missed because they are overshadowed by "neutral."

Figure 1-Right presents the dominant emotion profiles for all participants after excluding "neutral," allowing us to see more nuanced and unique emotion profiles. Despite the unique distribution of dominant emotions per participant, we find some commonalities: (1) emotions with a negative valence (i.e., anger, contempt, disgust, fear, and sadness) appear to be dominant more often than emotions with a positive valence (i.e., happiness and surprise); and (2) most people exhibit anger, happiness, sadness, and surprise as their dominant emotion at least once during the day.

Clustering participants by gender, we find that women exhibit what the tool classifies as sadness as the most dominant emotion, whereas for men it's a combination of positive and negative emotions including anger, sadness, happiness, and surprise. We choose only these two gender groups for comparison given our participants' self-reported gender identification on an open-text response. Figure 1 also shows the dominant emotion profiles for the average female and male in our participant pool. Finding that men express more anger than women is consistent with prior socio-cultural work [7, 35]. However, women expressing a greater proportion of sadness than men comes as a surprise: prior work finds women to display more positive emotions than men [7]. We consider this in more detail in the upcoming subsections.

4.2.2 Emotion Spikes. While dominant emotion as a metric highlights the most prominent emotions exhibited by our participants that are recorded by the AER tool, it overlooks the sensitivity of emotions that might not be the most prominent but show higher values than their usual baseline. For example, contempt and disgust are rarely observed as dominant emotions, but this does not mean that their magnitudes are meaningless. If we normalize and plot each emotion per day, we can identify "spikes" in even uncommon

emotions: times when their values were higher during the day, calculated relative to their mean value. We confirm that these spikes in emotions are not random noise by only including emotions if their magnitude is ≥ 0.20 on a scale of 0–1, same as the probability value for each emotion. Our threshold value of 0.20 was selected based on a grid search, optimizing on a good signal-to-noise ratio, and was validated with feedback from the interviews.

Analyzing data on emotion spikes per participant (Figure 2), we find that, once again, people have fairly unique profiles. Even though sadness was the most dominant emotion on average for women, and happiness was one of the dominant emotions for men, women have more spikes for happiness throughout the day than men. Women also tend to display overall more spikes in various emotions. Calculating the percentage of times each emotion spikes during the day per participant, we see that women show more emotion spikes throughout the day as compared to men (average values for all emotions except for fear are equal or higher for women).

4.2.3 Baseline Emotions. So far, we have seen conflicting results from dominant and spiking emotions: happiness is a dominant emotion for men instead of women, but women have higher percentages of happiness spikes throughout their day than men. We hypothesize that this is due to sadness being a *baseline* observed emotion for women, i.e., women tend to exhibit more expressions of what the tool identifies as sadness when working on a computer. To corroborate this, we calculate "baseline emotions" for all participants: any emotion that is observed at a non-zero magnitude for 90% of the day is tagged as a baseline emotion. These are different from both dominant emotions (the emotions with the highest magnitude in a given timeframe) and emotion spikes (the emotions that have a value higher than their mean in a given timeframe). From this point forward, we study the observed emotion profile for emotion spikes after removing any baseline emotions from it.

Indeed, we find that all women participants have sadness as one of their baseline emotions, whereas most men do not. While before, our results for women conflicted with prior work, after processing this emotion data using spiking and baseline emotions, the results are now consistent with a large body of prior literature on gender differences in the baserate expression of emotions [7, 75]. It is extremely encouraging that we can measure these differences in situ, and this is the first longitudinal evidence to support psychology theories of gender differences in the workplace context.

4.2.4 Summary of Observed Emotion. Overall, we see different patterns across participants for our various metrics. Dominant emotions represent the most prominent emotion label predicted for a participant's facial expressions for a given timeframe; emotion spikes reflect more sensitive and subtle emotion patterns, when an emotion is present more than its expected value; and baseline emotions represent the most common emotion labels predicted for an individual, prevalent throughout the day. While prior work only uses dominant emotion to describe people's affective state, we find that using dominant emotions (after removing "neutral"), emotion spikes, and baseline emotions is necessary to adequately capture observed emotion. If we were to only use the signal from dominant emotions, affective computing applications would not register any differences in people's affective states in the workplace. Additionally, with our new metrics (emotion spikes and baseline emotions),

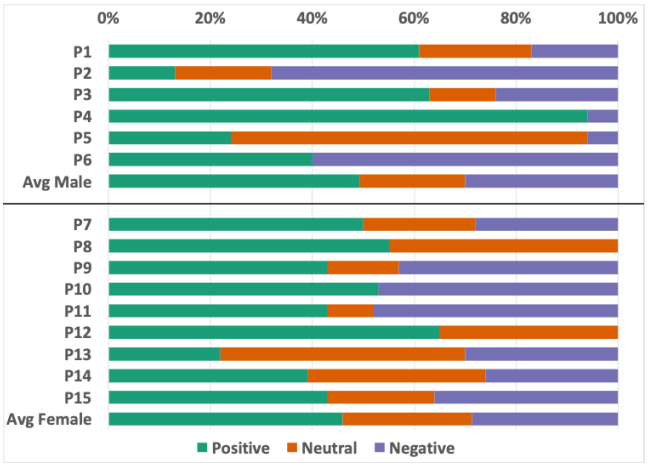

Figure 3: Reported affect from people's diaries, categorized as having positive, neutral, or negative valence.

we see consistency in our work and psychology studies of emotion baserates for different genders. We hypothesize that the workplace setting is the key differentiator: while dominant emotions alone might explain people's affective states in face-to-face interaction, the other metrics capture emotion in a context where people are not naturally expressive (here, in front of a computer screen).

4.3 Understanding Reported Affect States

We now look at self-reported affect values recorded by participants in their diaries (Table 1). The valence of the reported affect values (Figure 3) shows quite different patterns from those of dominant observed emotions seen above (comparing Figures 1 and 3). Contrary to the dominant emotion numbers, we find our reported affect sample to have more distributed valence, i.e., a balance of positive, negative, and neutral affect during their day, with no valence being significantly more present than others. These results are more aligned with the emotion spikes data obtained using the AER tool which shows a more subtle profile of emotions throughout the day (Figure 2). Once again, this corroborates our claim that emotion spikes and baseline emotions represent user affect via observed emotion labels more accurately than dominant emotions.

4.4 Comparing Observed Emotion and Reported Affect

So far, we have used descriptive statistics to identify patterns in observed emotions (dominant, spiking, and baseline) and self-reported user affect. Next, we compare dominant emotions and emotion spikes (after removing baseline emotions) with reported affect.

4.4.1 Dominant Observed Emotion vs. Reported Affect. A chi-square test between dominant observed emotions and reported affect shows significant difference between the valence of each ($\chi^2(2, N=215) = 192.78$, $p << 0.0001$, $V = 0.47$). This means that the observed emotion measured via the dominant emotion metric does not correspond well with the reported affect. We find the accuracy—calculated as the number of data points that have the same valence for dominant emotion and reported affect—to be 35.4%; Table 2-Left shows the distribution of classifications for each valence category.

AER Tool – Dominant Emotion

Diary	Positive	Neutral	Negative	Total
Positive	**5**	33	41	79
Neutral	13	**41**	23	77
Negative	2	27	**30**	59
Total	20	101	94	215

AER Tool – Emotion Spikes

Diary	Positive	Neutral	Negative	Total
Positive	**12**	41	26	79
Neutral	11	**49**	17	77
Negative	11	31	**17**	59
Total	34	121	60	215

Table 2: Classification values for valence of observed emotion (AER tool) and reported affect (diary). Left: Contingency matrix that uses dominant emotions, Right: Contingency matrix that uses emotion spikes (without baseline emotions).

The biggest error categories are when positive valence reported affect from the diary is classified as neutral or negative by the AER tool (Table 2-Left). This class of error is not surprising given our knowledge of dominant emotions, which tend to be negative (Figure 1), and thus would lead to most classifications being negative. When going over their diary and log data during the follow-up interview, this was a common type of error corrected by participants.

4.4.2 Observed Emotion Spikes vs. Reported Affect. Even when comparing data from emotion spikes (after removing baseline emotions), the chi-squared numbers continue to show significant difference (χ^2(2, N=215) = 75.58, $p << 0.0001$, $V = 0.30$) and the accuracy remains low (36.3%). It is important to note here that the value of the χ^2 statistic does improve: the new value is 75.58 compared to 192.78 before. Lower χ^2 statistic values indicate a reduced gap between the observed and the expected values, thus tending towards more alignment in this case. Analyzing the type of errors, we find that while neutral and positive valence of affect showed improvement, negative valence classifications got worse (Table 2-Right).

4.4.3 Summary of Comparison. Our results indicate that neither dominant nor spiking emotions perform well in terms of alignment with self-reported affect data from people's diaries, despite the fact that the AER tool we used is considered state-of-the-art in its ability to classify emotion. However, we cannot study workplace affect in isolation: people's affective states at work are greatly influenced by their context [49]. The AER tool records rich workplace context data; next, we look at how this might be incorporated to better understand and potentially align these signals. We do so to assess if it is possible to achieve convergent validity among these signals.

4.5 Factoring Workplace Context into Observed Emotion Profiles

We update people's observed emotion profiles using workplace context captured by the AER tool for two reasons. First, from the diary entries, we observed that people's tasks and workplace activity are related to their affective state. Participants often recorded the same emotion for different activities or different emotions for similar activities. For example, one of the tasks P13 noted in her diary was "get agent features from Q-learning," and mentioned "feeling good" about it at 3pm, but after working on it for a while, at 3:45pm, noted that she was "feeling eh, this is slow and tiring." Second, from people's subjective evaluations of the AER tool during the follow-up interviews, we noted that a majority of them considered the workplace context captured by AER tool to be a good reflection on the tasks they were doing throughout the day. For example, P9 indicated that the workplace activity data from the AER tool accurately represented their to-do list for the day: "*I could tell you my to-dos from yesterday from just looking at the activity data [mouse and keyboard activity rates]—that's kinda amazing. I guess I didn't expect it would be so accurate.*"

The AER and context tool logs workplace context data in the background, including mouse and keyboard activity (not recording exact key strokes or what was clicked, just the number), number of active windows and tabs (no titles recorded to preserve privacy), distance from screen, and eye movement. We use this additional context to define an *active* state, when mouse or keyboard activity is high (i.e., someone is actively writing) and distance from screen is low (leaning in). The thresholds for these states were determined on an individual basis: when the usage rates for mouse and keyboard, tab and window switches, etc., were higher than the daily average; when the distance from the screen was lower than the daily average. From our observations on baseline emotions and from follow-up interviews with people, we know that the AER tool assigns negative emotions for most people when they are deeply engaged in a task, even though, in practice, these states are considered positive by people. We use this knowledge to update the valence of dominant and spiked emotions to be positive when we observe negative or neutral emotion while being in an active state (this is effectively similar to coding "determined" as positive in the diary entries).

Using this new valence that takes workplace context into consideration, we are able to achieve 58.6% accuracy in aligned classifications between spiked emotions (– baseline emotions + activity) and reported affect from the diaries (22.3% absolute and 61.5% relative improvement in accuracy). Figure 4 shows participants' valence profiles based on this updated emotion spikes data from the AER tool and the reported affect data from the diaries, using only the 215 data points for which both types of affect data is available. The two sources are visibly more aligned for a majority of the participants, and a chi-squared test reveals the same. We find no significant difference between the two sources of valence (χ^2(2, N=215) = 1.86, $p = 0.39$, $V = 0.05$), but note some remaining sources of error in the updated contingency matrix (Table 3-Right). Compared to before, the inclusion of workplace activity data improves the positive and negative classifications made by the AER tool—only 8 positive and 7 negative reported affect time intervals are classified with the opposite valence by the AER tool. However, there is an increased misalignment for neutral valence. We note that the workplace activity data adds a positive bias to the neutral classifications (28 compared to 11 before). This is not unexpected given the way in which workplace activity is incorporated. For dominant emotions,

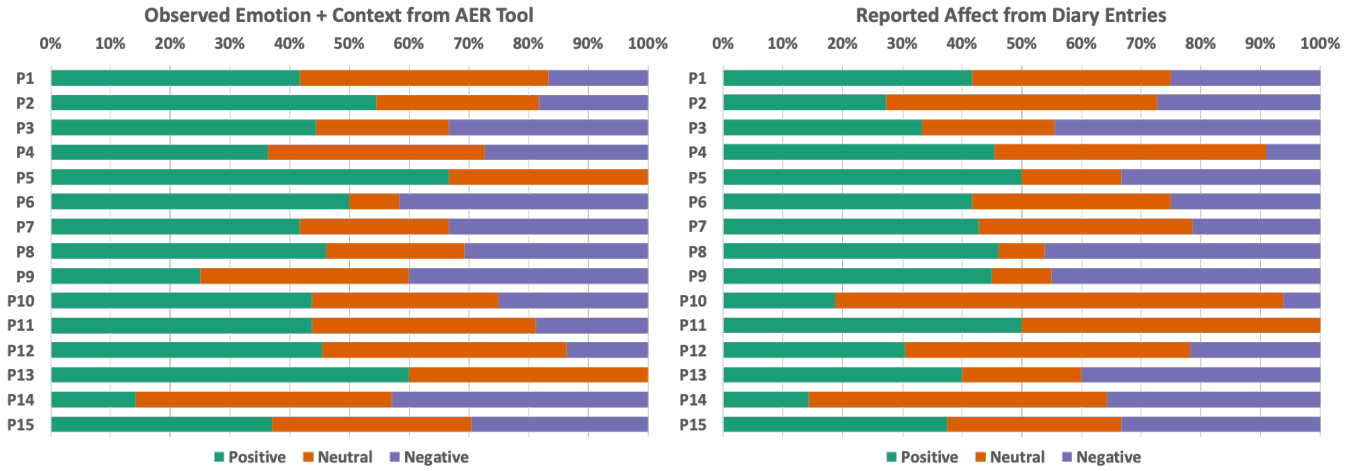

Figure 4: Emotion valence profiles for all participants using the 215 data points for which both types of affect data is available. Left: Valence profile based on the best performing observed emotion and context data from AER tool, i.e., emotion spikes (– baseline emotions + activity data). Right: Valence profile using reported affect from diaries.

AER Tool – Dominant Emotion + Workplace Activity

Diary \ AER Tool	Positive	Neutral	Negative	Total
Positive	**15**	43	21	79
Neutral	24	**43**	10	77
Negative	32	20	**7**	59
Total	71	106	38	215

AER Tool – Emotion Spikes + Workplace Activity

Diary \ AER Tool	Positive	Neutral	Negative	Total
Positive	**53**	18	8	79
Neutral	28	**36**	13	77
Negative	7	15	**37**	59
Total	88	69	58	215

Table 3: Classification values for valence of observed emotion (AER tool) and reported affect (diary), updated using workplace activity data. Left: Updated contingency matrix that uses dominant emotions (+ activity data). Right: Updated contingency matrix that uses spiked emotions (– baseline emotions + activity data).

the activity-based change in valence decreases the accuracy of classifications to 30.2%, thus increasing misalignment (Table 3-Left). A chi-square test continues to show significant differences between the valence of dominant emotions (+ activity) and reported affect ($\chi^2(2, N=215) = 20.44, p << 0.0001, V = 0.15$).

Overall, emotion spikes continue to perform better than dominant emotions in representing people's affective states in the workplace: the former shows 28.4% absolute and 94% relative improvement when compared to dominant emotions (+ activity). Hypothesizing about the underlying reason for these comparison statistics, we believe that emotion spikes are a better representation of observed emotion to align with self-report data. Emotion spikes capture the sensitivity of emotion categories, i.e., when an emotion is not dominant throughout, but shows an increased presence compared to its usual baseline. At the heart of it, reported affect represents a similar affect profile: people naturally remember the emotions that represent the highlights of a time interval and note these in their diaries. These "spotlight moments" appear to be better aligned with emotion spikes, at least in our results. We hope future work can further disentangle the relationships between these affect profiles derived using various metrics.

4.6 Summary of Results

Our study of observed emotions from an AER and context logging tool and reported affect from self-reports in a diary revealed two key aspects of capturing affect in the workplace. *First*, we identified a need for updating the existing dominant emotion metric used to capture emotion labels from AER tools, and suggested two new metrics for capturing observed emotion specifically for the workplace context: emotion spikes and baseline emotions. Dominant emotions do not reflect the subtleties of emotion expression in a setting where people do not naturally emote (here, in front of a screen). The new metrics suggested based on our study were more effective in capturing observed emotion in the workplace. *Second*, a comparison between the two sources of affect reveals misalignment to be a significant issue. While this misalignment is improved when we include workplace context to update the observed emotion signals, the maximum alignment we achieve is 58.6%. Once again, observed emotion data from emotion spikes (without baseline emotions, with workplace context) is a better signal than dominant emotions, but the continued misalignment must be taken into consideration as we think about applying these AER tools to obtain a continuous signal for affective computing applications.

5 DISCUSSION AND FUTURE WORK

The heart of affective computing is the intersection of human emotions with common computing tasks. Given how personal emotions can be to people, especially in workplaces where they may feel vulnerable, it is important that our measurement and treatment of affect be precise. If AER tool outputs were to be used without further processing, we would not capture any useful emotion signal for the workplace, as seen in our results for dominant emotions. Our new metrics for observed emotion allow for a more nuanced comparison between the two signals we study, though these metrics continue to be somewhat misaligned with self-reported affect. Next, we discuss potential sources for this mismatch between observed emotion and reported affect, the impact of this mismatch on the feasibility of using AER tools in practice, and describe the challenges of construct validity for workplace affect profiles.

5.1 Observed Emotion and Reported Affect are Different Signals

We hypothesize two potential reasons for the mismatch between our two affect signals: (1) external emotion labels and internal self-reports are shaped by context in different ways—on the surface, they appear misaligned, even though they are being influenced by the same context; (2) differences occur because general facial expression models are not appropriate for all interaction contexts—talking face-to-face with a person yields different facial expressions than when working in front of a computer screen (as anticipated by [3, 64] and others). Models evaluated for face-to-face interaction may not transfer to the nuances of a new setting, and workplace has an abundance of these.

With our combined tool-use and diary study followed by interviews, conducted in a constrained workplace context, we show concrete evidence for Barrett et al.'s claim: you cannot evaluate AER technology in isolation [3]. We also further qualify this claim: instead of trying to align these automated and self-reported affect signals, we propose thinking of them as different signals. Observed emotions are likely capturing the expression people project, and this projection is shaped by different external contexts (e.g., looking at a screen at work). Thus, observed emotions represent, to some capacity, the affect profile based on external contexts (i.e., physical and temporal aspects of an individual's surroundings). Reported affect, on the other hand, reflects how people are internally feeling, and thus represents, to some capacity, the affect profile based on internal contexts (i.e., history or current state, specific to an individual). Given this framework of interpretation for these signals, we can confidently say that: (1) affect must be studied in the context it is being expressed, and (2) neither signal should replace the other for affective computing applications.

5.2 The Challenge of Reverse Inference in Emotion Expression

Supporting prior work's critiques, our study shows concrete evidence of a lack of one-to-one mapping between facial expressions and emotions labels, in this new workplace context. Even if facial movement can be reliably and specifically captured, it is the reverse inference to an emotion expression that is problematic. Indeed, both emotion expression and emotion perception are contextual tasks, dependent on the person, the setting, temporal aspects, etc. [3, 8, 64]. We observed similar challenges in our study of AER technology: while the tool captured important and nuanced signals, it consistently mislabeled some of the facial movements (e.g., yawning was labeled as "surprise" for P15). Though this mislabeling is problematic, the consistency with which it occurred within a particular context indicates that there are opportunities for obtaining accurate emotion labels for these facial movements. One of our participants suggested that they might want to personalize their emotion labels anyway—this could help correct the labels and provide personalized, contextual emotion expression labeling that would resolve some of the critiques raised by prior work. Our hope is that future work will conduct contextual studies similar to ours to verify the reliability and specificity of AER technology before they are used in real-world applications.

5.3 Reported Affect Is Not the Ground Truth

Attempting to align observed emotion from AER tools to self-reported affect might suggest that reported affect is the ground truth, but that would be an unfair assessment. People are also biased in providing data about their feelings: when asked about several surprise spikes in their data at the end of the day, P12 mentioned that they *"found this really awesome paper in my lit search that was so applicable to my work, I was like holy shit this changes everything! I totally forgot that I found that paper by the way, so thanks for the reminder. Of course also forgot to note it [in the diary]. Whoops!"*. Similarly, P13 had happiness spikes right after a meeting, which we assumed were from checking social media, but: *"On most days that might have been a yes, but yesterday I just had an awesome meeting and I remember being very happy after it. I guess I forgot to note that for that diary interval."* Not only can people be inconsistent in their reporting, they also sometimes find it hard to describe how they are feeling. Given these idiosyncrasies and biases in self-reported affect, we should not think of it as the ground truth, but rather, a behavioral signal for people's affect. Prior work also suggests this as the most reasonable use of self-reported signals [48, 141].

5.4 Designing Adequate Measures of Affect

Given the uniquely different benefits and challenges of recording and using observed emotion and reported affect signals, finding perfect alignment between the two should not be the goal. AER technology and self-reports both capture important contextual information and have different advantages. AER technology is able to consistently capture facial landmarks, even if it cannot yet assign accurate emotional inferences to the different facial configurations that are captured. It provides a continuous affect signal; does not require constant, direct input from users; and can be applied at a large scale. Self-reports of affect can be more representative of how people feel at discrete moments. Instead of replacing either, we propose evaluating whether the varied affect signals are capturing useful information and, if so, designing ways to combine these to create a more representative affect profile. HCI research has shown evidence of successful implementations of this approach in the form of mixed-initiative interfaces [56]: when the goals and needs of users are uncertain or hard to capture, we can design opportunities for users and automated approaches to collaborate

for better outcomes. This type of conditional delegation has shown some success for content moderation [10] and fake speech detection [114], both of which are domains with similar challenges of ground truth. A similar approach for affect signals could be designed by asking people for periodic feedback in conjunction with AER outputs (e.g., [67]) or human-in-the-loop observed emotion labeling (e.g., via continuous annotation [139]). We are excited to see other innovative approaches for this task in future work.

5.5 AER Technology in Remote Work Settings

With the rising prevalence of remote work in the workplace context, AER technology can be a helpful resource for affective computing applications at the individual, group, and organization levels. Remote work is by no means a new setting. Olson and Olson [100], and others (e.g., [65, 71, 101]) have extensively studied remote work and how to navigate its collaboration challenges (e.g., when work in a remote team is tightly-coupled). Below, we highlight some applications of AER technology for supporting remote work.

For individuals, attention management is even more important with remote work. Monitoring emotional valence and arousal such that people are not interrupted with notifications or online events during their productive times could be made possible with AER technology in combination with machine learning models. Additionally, as mentioned by our participants, the output visualizations based on data from AER technology can be used for self-tracking, work journaling, and keeping track of both mood and productivity at work. Prior work has shown that happiness is central to unlocking productivity gains [47, 49], and AER technology can help guide people to specific emotional states [67, 70]. At the group level, AER technology can be particularly useful for accessibility-related affective computing applications (e.g., for autism [143], visual impairments [119]). People with disabilities can be guided through a group meeting or collaborative, remote work settings using outputs from AER technology from others in the room, shared with their consent. Another application can target remote meetings with a large number of participants: monitoring emotional valence and arousal for the purposes of deciding turn-taking in conversations, to ensure that everyone in the "room" can present their views if they so desire. These tools can also provide feedback to people presenting in a remote meeting, to give them a better sense of the "room"'s reaction to their presentation and if clarifications are needed at any point. At the organization level, AER can help with measurements of job-related stress, and identifying and understanding factors causing this stress. These measurements can subsequently provide initial guidance on how these stressors could be addressed [135]. With all these applications, there are privacy considerations to bear in mind: data sharing, storage, and access will all be important avenues for future work as a part of developing these applications.

5.6 Ethical Considerations

A critical ethical consideration with AER is the potential for using the underlying audio and video data for unscrupulous applications. Similar to facial recognition—where facial analysis is used to identify a particular individual in privacy-invading settings [45, 61]—one could imagine the audio and video needed by AER being applied for identification and surveillance [144]. The tool that we selected for our study circumvented this by never storing this data and predicting output emotion labels in real time instead. This required significant computational resources and restricted our study setup to people's workstation desktops, but it also made them feel more comfortable about participating in our study. We hope that future work will continue this line of work to make AER privacy-preserving in more (computational) resource-friendly ways, and follow ethical guidelines for designing emotion recognition systems [53].

However, this does not mean that using our type of AER and context logging tool is without privacy and bias concerns. The underlying datasets for AER technology have similar biases to those used for facial recognition: they are not representative of diversity in populations in terms of, for example, race, gender, emotions, or ability [17, 68, 115]. Moreover, they are collected in a context that is not always the same as the context in which they are applied (i.e., posed, in-person emotion expressions vs., for example, emotion expression in front of a screen) [3, 64, 97]. Applying AER models built on these biased datasets can lead to harmful consequences for specific populations, especially if the output of the AER models are directly used in affective computing applications.

Without nuanced understanding of the misalignment between observed emotions and reported affect, AER outputs could also be misconstrued in different ways. As we found in our study, emotion labels might not match affect in the workplace context: "surprise" in front of a screen could be "tiredness" (as P15 notes) or "anger" could be "focus" (as is the case for P5). Our work shows that there is ambiguity around what these signals represent, and future work must address this before such applications can be considered in practice. For individual users, not doing so could result in inaccurate and unreliable signals being presented to users of this technology, and can harm their psyche. If used in organizational settings without any checks and balances, AER tools may lead to practices of monitoring people's affective states—a form of surveillance and a breach of privacy. Our participants noted these concerns and wanted autonomy over what signals were recorded, who would have access to the recorded data, and how it would be used. We agree that control of these affect-related outputs must remain with the individual, so that they can benefit from a better understanding of their affective states at work, rather than these outputs being misused as a surveillance mechanism.

5.7 Challenges with Construct Validity in Emotion Expression and Perception

We encountered two primary challenges related to construct validity in evaluating AER technology within the workplace context. First, the emotion expression setting has no ground truth. While prior work has often relied on self-reported affect values (e.g., using the PANAS scale) as the final word on people's affective states, we have shown here that reported affect is simply one signal of affect, and the variability and potential biases of reported affect do not lend themselves to its use as ground truth. Given this lack of a clear ground truth, a feasible path forward is to study convergent or discriminant construct validity with the affect signals available for use. Prior work in other settings that have similar challenges (e.g., determining mental health using information available online [9, 32]) has followed a similar methodology. We did

this for one AER technology here, which was selected for reasons described in Section 3.2. There are other affect signal alternatives to consider, for example, facial heat patterns [62], wearable sensors (e.g., smart eyewear [109]; EEG, EMG, and GSR sensors [44, 106]), fMRI scans [46], AER technology that relies on crowd- and expert-tagged facial [4, 129] and voice [94] markers, etc. A contextual evaluation of all these signals was infeasible and out of scope for this paper. We hope to see future work tackle this challenge in subsets or comprehensive setups. With more of these studies, we can better characterize the similarities and differences between these signals—a necessity for settings that lack ground truth.

The second challenge we faced was that emotion expression is dependent on the context in which the facial configurations are being captured and, unfortunately, operationalizing and capturing all relevant context is an infeasible task [124]. For example, consider the challenges of context that natural language processing is yet to find a solution for [42]. One way to circumvent this issue would be to rely on people to provide accurate annotations of their own data. Participants in our study were happy to do this when the task was not burdensome (e.g., finding that focus is always labeled as "anger" or yawning as "surprise", and correcting those labels). If we could find less cumbersome ways of obtaining more fine-grained, continuous feedback from people on these emotion and context outputs, we could vastly improve the accuracy of these tools in contextual settings. Recent work in HCI has shown promise in developing continuous annotation software (e.g., [43, 138, 139, 142]). We hope future work will continue improving these setups to help generate the kind of data we need for affect evaluation.

6 LIMITATIONS

Our study is limited by the technology and setup it relies on. We employed an existing AER tool [92] with state-of-the-art facial analysis and workplace context monitoring. We rely on prior work's claims for the performance of this tool and its facial analysis capabilities [4, 92]. An interesting avenue for future work would be a direct comparison between various AER tools and methodologies to verify if their outputs are aligned in a contextual setting. We compared the AER tool's output to reported affect collected via a diary study. We selected a diary given the opportunity for receiving nuanced subjective data for the same overall categories as the tool (affect and context) as well as maintaining participant privacy in data analysis. A limitation of this setup is that asking people to maintain a diary could be an interruption to their work. We tried to minimize this with our template (Section 3.3), but cannot be sure about the level of impact this had on our findings. Future work should consider other approaches (e.g., continuous emotion annotation of recorded video data from participants), while being careful of the data and participant privacy considerations of their setups.

There are also limitations to our analyses. Our data collection via the combined tool-use and diary study was done over the course of one day, with a follow-up interview conducted the next day. While this gave us sufficient data (total 104 hours of tool use data and 331 diary entries) for conducting the analyses that we have shown here, we cannot confirm how the resulting patterns might change over longer study periods. We believe our results would be consistent over time because of the diversity already captured within participant data, and corroboration for all data collection and analyses via the interview study, but future work could learn additional patterns via a longer, longitudinal study. Further, for our quantitative comparisons, we relied on valence as the high-level metric common to both signals. Other fine-grained metrics (e.g., basic emotions, continuously tagged subjective emotions) might also provide helpful insights on these signals. Relatedly, we encoded "surprise" as having a positive valence, even though prior work has been unclear on whether it is a positive or negative signal [99]. We did so because of evidence of its positive valence for this particular context based on some signals in prior work (e.g., satisfaction) as well as how our participants responded to it during the follow-up interviews. However, we acknowledge that this assumption might not hold true for larger populations or other contexts.

Finally, there are limitations related to our data. Most of our participants had research-focused roles. While the data showed diversity in both the emotions captured and the categories of tasks performed, and corroborated existing psychology theories around gender differences in emotion baserates, more/different emotion and task patterns could emerge from a representative sample of information workers. To that end, another limitation of our work is its comparison of only men and women in the gender categories [68]. Per existing gender reporting guidelines [116], we allowed participants to self-report their gender. We recognize that, as a result of this self-reported gender data, we cannot provide a complete understanding of gender identities with respect to this AER tool.

7 CONCLUSION

In this paper, we have presented results from an in-situ study comparing external *observed emotion* signal obtained via a state-of-the-art Automatic Emotion Recognition (AER) tool with internal *reported affect* states collected via self-reports in diaries. Via a day-long, combined tool-use and diary study ($N = 15$) we characterize different observed emotion profiles using dominant emotion signals, a common metric in previous affect sensing work, as well as emotion spikes and baseline emotions, two metrics that we develop to obtain more subtle workplace affect profiles. On comparing the two affect signals (observed emotion and reported affect), we find that they are misaligned (35.4% match), with alignment increasing up to 58.6% at most after including workplace activity and context data. Our results support a distinction between people's external-facing emotions and their internal affect states, although both signals can be messy to consistently capture. We discuss whether technology can—or even should—bridge this misalignment. Instead, we note aligning people's expectations of AER technology and identifying the common underlying emotion that AER's external and people's internal signal stem from as the more salient goals.

ACKNOWLEDGMENTS

We would like to thank our reviewers for their insightful critique, which helped improve the paper. We are also grateful to Mary Czerwinski, Michael Madaio, Stevie Chancellor, Eric Gilbert, Cliff Lampe, and the P+I and HUE teams at Microsoft Research for their support and feedback. This work was initiated while the first and third authors were interns at Microsoft Research.

REFERENCES

[1] Nalini Ambady and Robert Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin* 111, 2 (1992), 256.

[2] Brian P. Bailey and Joseph A. Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior* 22, 4 (1 7 2006), 685–708. https://doi.org/10.1016/j.chb.2005.12.009

[3] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* 20, 1 (2019), 1–68.

[4] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 279–283.

[5] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, Qianli Feng, Yan Wang, and Aleix M Martinez. 2017. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210* (2017).

[6] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological*. American Psychological Association, Washington, DC, US, 57–71. https://doi.org/10.1037/13620-004

[7] Leslie R Brody and Judith A Hall. 2008. Gender and emotion in context. *Handbook of emotions* 3 (2008), 395–408.

[8] John T Cacioppo and Louis G Tassinary. 1990. *Principles of psychophysiology: Physical, social, and inferential elements*. Cambridge University Press.

[9] Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1171–1184.

[10] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

[11] Chapter, Kim Cameron, Jane Dutton, and Robert Quinn. 2003. An Introduction to Positive Organizational Scholarship. (01 2003).

[12] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. 2003. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and image understanding* 91, 1-2 (2003), 160–187.

[13] Jeffrey F Cohn and Karen L Schmidt. 2004. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing* 2, 02 (2004), 121–132.

[14] Daniel T Cordaro, Rui Sun, Dacher Keltner, Shanmukh Kamble, Niranjan Huddar, and Galen McNeil. 2018. Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion* 18, 1 (2018), 75.

[15] Stanley Coren and James A Russell. 1992. The relative dominance of different facial expressions of emotion under conditions of perceptual ambiguity. *Cognition and Emotion* 6, 5 (1992), 339–356.

[16] Sarah Cosentino, Estelle IS Randria, Jia-Yeu Lin, Thomas Pellegrini, Salvatore Sessa, and Atsuo Takanishi. 2018. Group emotion recognition strategies for entertainment robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 813–818.

[17] Kate Crawford and Trevor Paglen. 2021. Excavating AI: The politics of images in machine learning training sets. *AI & SOCIETY* (2021), 1–12.

[18] Mary Czerwinski, Javier Hernandez, and Daniel McDuff. 2021. Building an AI That Feels: AI systems with emotional intelligence could learn faster and be more helpful. *IEEE Spectrum* 58, 5 (2021), 32–38.

[19] C Darwin. 1872. *The expression of emotions in man and animals* (3rd ed.). Oxford University, New York, NY.

[20] Richard J Davidson. 1999. Neuropsychological perspectives on affective styles and their cognitive consequences. (1999).

[21] Giuseppe Destefanis, Marco Ortu, Steve Counsell, Stephen Swift, Michele Marchesi, and Roberto Tonelli. 2016. Software development: do good manners matter? *PeerJ Computer Science* 2 (2016), e73.

[22] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. 2015. Recognition of genuine smiles. *IEEE Transactions on Multimedia* 17, 3 (2015), 279–294.

[23] Ulf Dimberg and Lars-Olov Lundquist. 1990. Gender differences in facial reactions to facial expressions. *Biological psychology* 30, 2 (1990), 151–159.

[24] Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 43.

[25] Juan I Durán, Rainer Reisenzein, and José-Miguel Fernández-Dols. 2017. Coherence between emotions and facial expressions. *The science of facial expression* (2017), 107–129.

[26] Paul Ekman. 1989. The argument and evidence about universals in facial expressions. *Handbook of social psychophysiology* 143 (1989), 164.

[27] Paul Ekman. 1992. Are there basic emotions? (1992).

[28] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[29] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).

[30] Paul Ekman, E Richard Sorenson, and Wallace V Friesen. 1969. Pan-cultural elements in facial displays of emotion. *Science* 164, 3875 (1969), 86–88.

[31] Hillary Anger Elfenbein. 2017. Emotional dialects in the language of emotion. *The science of facial expression* (2017), 479–496.

[32] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–16.

[33] Elizabeth C Evans. 1969. Physiognomics in the ancient world. *Transactions of the American Philosophical Society* 59, 5 (1969), 1–101.

[34] José-Miguel Fernández-Dols and Maria-Angeles Ruiz-Belda. 1997. Spontaneous facial behavior during intense emotional episodes: Artistic truth and optical truth. In *The psychology of facial expression*, James A Russell and José-Miguel Fernández-Dols (Eds.). Cambridge University Press, New York, NY, 255–274.

[35] Agneta H Fischer. 1993. Sex differences in emotionality: Fact or stereotype? *Feminism & Psychology* 3, 3 (1993), 303–318.

[36] Cynthia Fisher. 2000. Mood and emotions while working: Missing pieces of job satisfaction? *School of Business Discussion Papers* 21 (03 2000). https://doi.org/10.1002/(SICI)1099-1379(200003)21:2<185::AID-JOB34>3.0.CO;2-M

[37] Mark G Frank, Paul Ekman, and Wallace V Friesen. 1993. Behavioral markers and recognizability of the smile of enjoyment. *Journal of personality and social psychology* 64, 1 (1993), 83.

[38] Wallace V Friesen, Paul Ekman, et al. 1983. EMFACS-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco* 2, 36 (1983), 1.

[39] Anna-Katharina Frison, Philipp Wintersberger, Andreas Riener, Clemens Schartmüller, Linda Ng Boyle, Erika Miller, and Klemens Weigl. 2019. In UX We Trust: Investigation of Aesthetics and Usability of Driver-Vehicle Interfaces and Their Impact on the Perception of Automated Driving. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[40] Maria Gendron, Debi Roberson, Jacoba Marietta van der Vyver, and Lisa Feldman Barrett. 2014. Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. *Emotion* 14, 2 (2014), 251.

[41] Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. Towards understanding emotional intelligence for behavior change chatbots. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 8–14.

[42] Daniel Gildea. 2001. Corpus Variation and Parser Performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*. http://www.aclweb.org/anthology/W01-0521

[43] Jeffrey M Girard. 2014. CARMA: Software for continuous affect rating and media annotation. *Journal of open research software* 2, 1 (2014).

[44] Daniela Girardi, Filippo Lanubile, and Nicole Novielli. 2017. Emotion detection using noninvasive low cost sensors. In *2017 Seventh international conference on affective computing and intelligent interaction (ACII)*. IEEE, 125–130.

[45] Mitchell Gray. 2003. Urban Surveillance and Panopticism: will we recognize the facial recognition society? *Surveillance & Society* 1, 3 (2003), 314–330.

[46] Marcus A Gray, Neil A Harrison, Stefan Wiens, and Hugo D Critchley. 2007. Modulation of emotional appraisal by false physiological feedback during fMRI. *PLoS one* 2, 6 (2007), e546.

[47] Daniel Graziotin, Xiaofeng Wang, and Pekka Abrahamsson. 2014. Happy software developers solve problems better: psychological measurements in empirical software engineering. *PeerJ* 2 (2014), e289.

[48] James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology* 2, 3 (1998), 271–299.

[49] Barry Gruenberg. 1980. The happy worker: An analysis of educational and occupational differences in determinants of job satisfaction. *American journal of sociology* 86, 2 (1980), 247–271.

[50] Jamie Guillory, Jason Spiegel, Molly Drislane, Benjamin Weiss, Walter Donner, and Jeffrey Hancock. 2011. Upset now? Emotion contagion in distributed groups. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 745–748.

[51] Hatice Gunes and Maja Pantic. 2010. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions (IJSE)* 1, 1 (2010), 68–99.

[52] Judith A Hall and David Matsumoto. 2004. Gender differences in judgments of multiple emotions from facial expressions. *Emotion* 4, 2 (2004), 201.

[53] Javier Hernandez, Josh Lovejoy, Daniel McDuff, Jina Suh, Tim O'Brien, Arathi Sethumadhavan, Gretchen Greene, Rosalind Picard, and Mary Czerwinski. 2021. Guidelines for Assessing and Minimizing Risks of Emotion Recognition Applications. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[54] Ursula Hess, Arvid Kappas, Gregory J McHugo, Robert E Kleck, and John T Lanzetta. 1989. An analysis of the encoding and decoding of spontaneous and posed smiles: The use of facial electromyography. *Journal of Nonverbal Behavior* 13, 2 (1989), 121–137.

[55] Jesse Hoey. 2001. Clustering facial displays in context. *Technical Report TR-01-17* (2001).

[56] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[57] Shamsi Iqbal, Jina Suh, Mary Czerwinski, Gloria Mark, and Jaime Teevan. 2020. Remote Work and Well-being. The New Future of Work Symposium, Published by Microsoft. https://www.microsoft.com/en-us/research/publication/remote-work-and-well-being/

[58] Shamsi T. Iqbal and Eric Horvitz. 2007. Disruption and Recovery of Computing Tasks: Field Study, Analysis, and Directions. In *In Proceedings of the Conference on Human Factors in Computing Systems - CHI 2007 (Apr. 28-May 3)*. ACM, 677–686.

[59] Carroll E Izard. 1994. Innate and universal facial expressions: evidence from developmental and cross-cultural research. (1994).

[60] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. 2012. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences* 109, 19 (2012), 7241–7244.

[61] Anil K Jain and Stan Z Li. 2011. *Handbook of face recognition*. Springer.

[62] Sophie Jarlier, Didier Grandjean, Sylvain Delplanque, Karim N'diaye, Isabelle Cayeux, Maria Inés Velazco, David Sander, Patrik Vuilleumier, and Klaus R Scherer. 2011. Thermal analysis of facial muscles contractions. *IEEE transactions on affective computing* 2, 1 (2011), 2–9.

[63] Ashish Kapoor and Rosalind W Picard. 2005. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia*. ACM, 677–682.

[64] Arvid Kappas. 2003. What facial activity can and cannot tell us about emotions. In *The human face*. Springer, 215–234.

[65] Demetrios Karis, Daniel Wildman, and Amir Mané. 2016. Improving remote collaboration with video conferencing and video portals. *Human–Computer Interaction* 31, 1 (2016), 1–58.

[66] Charalampos Karyotis, Faiyaz Doctor, Rahat Iqbal, Anne James, and Victor Chang. 2016. A Fuzzy Modelling Approach of Emotion for Affective Computing Systems. (2016).

[67] Harmanpreet Kaur, Alex C Williams, Daniel McDuff, Mary Czerwinski, Jaime Teevan, and Shamsi T Iqbal. 2020. Optimizing for Happiness and Productivity: Modeling Opportune Moments for Transitions and Breaks at Work. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[68] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 88.

[69] Latika Kharb and Sarabjit Kaur. [n.d.]. Embedding Intelligence through Cognitive Services. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN* ([n. d.]), 2321–9653.

[70] Everlyne Kimani, Kael Rowan, Daniel McDuff, Mary Czerwinski, and Gloria Mark. 2019. A Conversational Agent in Support of Productivity and Wellbeing at Work. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–7.

[71] Steve WJ Kozlowski and Daniel R Ilgen. 2006. Enhancing the effectiveness of work groups and teams. *Psychological science in the public interest* 7, 3 (2006), 77–124.

[72] Eva G Krumhuber and Antony SR Manstead. 2009. Can Duchenne smiles be feigned? New evidence on felt and false smiles. *Emotion* 9, 6 (2009), 807.

[73] Theodore Kunin. 1955. The Construction of a New Type of Attitude Measure. *Personnel Psychology* 8, 1 (1955), 65–77. https://doi.org/10.1111/j.1744-6570.1955.tb01189.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-6570.1955.tb01189.x

[74] Marianne LaFrance and Mahzarin Banaji. 1992. Toward a reconsideration of the gender-emotion relationship. *Emotion and social behavior* 14 (1992), 178–201.

[75] Marianne LaFrance, Marvin A Hecht, and Elizabeth Levy Paluck. 2003. The contingent smile: A meta-analysis of sex differences in smiling. *Psychological bulletin* 129, 2 (2003), 305.

[76] Hung-Te Lee, Rung-Ching Chen, and David Wei. 2017. Building emotion recognition control system using Raspberry Pi. In *International Conference on Frontier Computing*. Springer, 36–45.

[77] Madelene Lindström, Anna Ståhl, Kristina Höök, Petra Sundström, Jarmo Laaksolathi, Marco Combetto, Alex Taylor, and Roberto Bresin. 2006. Affective diary: designing for bodily expressiveness and self-reflection. In *CHI'06 extended abstracts on Human factors in computing systems*. 1037–1042.

[78] Andrej Luneski, Panagiotis D Bamidis, and Madga Hitoglou-Antoniadou. 2008. Affective computing and medical informatics: state of the art in emotion-aware medical applications. *Studies in health technology and informatics* 136 (2008), 517.

[79] Fred Luthans. 2002. The need for and meaning of positive organizational behavior. *Journal of Organizational Behavior* 23, 6 (2002), 695–706. https://doi.org/10.1002/job.165 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/job.165

[80] Diana MacLean, Asta Roseway, and Mary Czerwinski. 2013. MoodWings: a wearable biofeedback device for real-time stress intervention. In *Proceedings of the 6th international conference on PErvasive Technologies Related to Assistive Environments*. ACM, 66.

[81] Muharram Mansoorizadeh and Nasrollah Moghaddam Charkari. 2010. Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications* 49, 2 (2010), 277–297.

[82] Mika Mäntylä, Bram Adams, Giuseppe Destefanis, Daniel Graziotin, and Marco Ortu. 2016. Mining valence, arousal, and dominance: possibilities for detecting burnout and productivity?. In *Proceedings of the 13th International Conference on Mining Software Repositories*. ACM, 247–258.

[83] Gloria Mark, Mary Czerwinski, Shamsi Iqbal, and Paul Johns. 2016. Workplace Indicators of Mood: Behavioral and Cognitive Correlates of Mood Among Information Workers. In *Proceedings of the 6th International Conference on Digital Health Conference*. ACM, 29–36.

[84] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. 2014. Capturing the mood: facebook and face-to-face encounters in the workplace. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1082–1094.

[85] Gloria Mark, Shamsi Iqbal, Mary Czerwinski, and Paul Johns. 2015. Focused, Aroused, but So Distractible: Temporal Perspectives on Multitasking and Communications. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. ACM, New York, NY, USA, 903–916.

[86] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, and Paul Johns. 2014. Bored mondays and focused afternoons: the rhythm of attention and online activity in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3025–3034.

[87] Brais Martinez and Michel F Valstar. 2016. Advances, challenges, and opportunities in automatic facial expression recognition. In *Advances in face detection and facial image analysis*. Springer, 63–100.

[88] David Matsumoto. 1990. Cultural similarities and differences in display rules. *Motivation and Emotion* 14, 3 (1990), 195–214.

[89] Daniel McDuff, Eunice Jun, Kael Rowan, and Mary Czerwinski. 2019. Longitudinal Observational Evidence of the Impact of Emotion Regulation Strategies on Affective Expression. *IEEE Transactions on Affective Computing* (2019).

[90] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. AffectAura: an intelligent system for emotional memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 849–858.

[91] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. 2016. AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*. 3723–3726.

[92] Daniel McDuff, Kael Rowan, Piali Choudhury, Jessica Wolk, ThuVan Pham, and Mary Czerwinski. 2019. A Multimodal Emotion Sensing Platform for Building Emotion-Aware Applications. *arXiv preprint arXiv:1903.12133* (2019).

[93] Daniel S Messinger, Leticia Lobo Duvivier, Zachary E Warren, Mohammad Mahoor, Jason Baker, Anne Warlaumont, and Paul Ruvolo. 2015. Affective computing, emotional development, and autism. (2015).

[94] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. 2008. Audio-visual emotion recognition using gaussian mixture models for face and voice. In *2008 Tenth IEEE International Symposium on Multimedia*. IEEE, 250–257.

[95] Sebastian C Müller and Thomas Fritz. 2015. Stuck and frustrated or in flow and happy: sensing developers' emotions and progress. In *Software Engineering (ICSE), 2015 IEEE/ACM 37th IEEE International Conference on*, Vol. 1. IEEE, 688–699.

[96] Prasanth Murali, Javier Hernandez, Daniel McDuff, Kael Rowan, Jina Suh, and Mary Czerwinski. 2021. AffectiveSpotlight: Facilitating the Communication of Affective Responses from Audience Members during Online Presentations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[97] Shushi Namba, Russell S Kabir, Makoto Miyatani, and Takashi Nakao. 2018. Dynamic displays enhance the ability to discriminate genuine and posed facial expressions of emotion. *Frontiers in psychology* 9 (2018), 672.

[98] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing* 2, 2 (2011), 92–105.

[99] Marret K Noordewier and Seger M Breugelmans. 2013. On the valence of surprise. *Cognition & emotion* 27, 7 (2013), 1326–1334.

[100] Gary M Olson and Judith S Olson. 2000. Distance matters. *Human–computer interaction* 15, 2-3 (2000), 139–178.

[101] Judith S Olson, E Hofer, Nathan Bos, Ann Zimmerman, Gary M Olson, Daniel Cooney, and Ixchel Faniel. 2008. A theory of remote scientific collaboration. *Scientific collaboration on the internet* (2008), 73–97.

[102] Marco Ortu, Bram Adams, Giuseppe Destefanis, Parastou Tourani, Michele Marchesi, and Roberto Tonelli. 2015. Are bullies more productive?: empirical study of affectiveness vs. issue fixing time. In *Proceedings of the 12th Working Conference on Mining Software Repositories*. IEEE Press, 303–313.

[103] H Oster, L Daily, and P Goldenthal. 1989. Processing facial affect. In *Handbook of research on face processing*, A W Young and H D Ellis (Eds.). Elsevier, Amsterdam, 101–161.

[104] Talcott Parsons, Robert Freed Bales, and James Olds. 1956. *Family socialization and interaction process*. Psychology Press.

[105] Monica Perusquía-Hernández. 2021. Are people happy when they smile?: Affective assessments based on automatic smile genuineness identification. *Emotion Studies* 6, 1 (2021), 57–71.

[106] Monica Perusquía-Hernández, Saho Ayabe-Kanamura, Kenji Suzuki, and Shiro Kumano. 2019. The invisible potential of facial electromyography: a comparison of EMG and Computer Vision when distinguishing posed from spontaneous smiles. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–9.

[107] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17, 3 (2005), 715–734.

[108] James A Russell. 1994. Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies. *Psychological Bulletin* 115, 1 (1994), 102.

[109] Chisa Saito, Katsutoshi Masai, and Maki Sugimoto. 2020. Classification of spontaneous and posed smiles by photo-reflective sensors embedded with smart eyewear. In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction*. 45–52.

[110] Najmeh Samadiani, Guangyan Huang, Borui Cai, Wei Luo, Chi-Hung Chi, Yong Xiang, and Jing He. 2019. A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors* 19, 8 (2019), 1863.

[111] Samiha Samrose, Wenyi Chu, Carolina He, Yuebai Gao, Syeda Sarah Shahrin, Zhen Bai, and Mohammed Ehsan Hoque. 2019. Visual Cues for Disrespectful Conversation Analysis. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 580–586.

[112] Samiha Samrose, Daniel McDuff, Robert Sim, Jina Suh, Kael Rowan, Javier Hernandez, Sean Rintel, Kevin Moynihan, and Mary Czerwinski. 2021. MeetingCoach: An Intelligent Dashboard for Supporting Effective & Inclusive Meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[113] Samiha Samrose, Ru Zhao, Jeffery White, Vivian Li, Luis Nova, Yichen Lu, Mohammad Rafayet Ali, and Mohammed Ehsan Hoque. 2018. Coco: Collaboration coach for understanding team dynamics during video conferencing. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 4 (2018), 1–24.

[114] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.

[115] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[116] Morgan Klaus Scheuerman, Katta Spiel, Oliver L. Haimson, Foad Hamidi, and Stacy M. Branham. [n.d.]. HCI Guidelines for Gender Equity and Inclusivity. ([n. d.]). https://docs.wixstatic.com/ugd/eb2cd9_3d08c68141ea4900b82b47a551a5c95f.pdf

[117] Karen L Schmidt, Zara Ambadar, Jeffrey F Cohn, and L Ian Reed. 2006. Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling. *Journal of nonverbal behavior* 30, 1 (2006), 37–52.

[118] Garima Sharma and Abhinav Dhall. 2021. *A Survey on Automatic Multimodal Emotion Recognition in the Wild*. Springer International Publishing, Cham, 35–64. https://doi.org/10.1007/978-3-030-51870-7_3

[119] Lei Shi, Brianna J Tomlinson, John Tang, Edward Cutrell, Daniel McDuff, Gina Venolia, Paul Johns, and Kael Rowan. 2019. Accessible Video Calling: Enabling Nonvisual Perception of Visual Conversation Cues. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–22.

[120] Stephanie A Shields, Dallas N Garner, Brooke Di Leone, and Alena M Hadley. 2006. Gender and emotion. In *Handbook of the sociology of emotions*. Springer, 63–83.

[121] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.

[122] Ruth Stock, Moritz Merkle, Dietmar Eidens, Martin Hannig, Paul Heineck, Mai Anh Nguyen, and Johannes Völker. 2019. When Robots Enter Our Workplace: Understanding Employee Trust in Assistive Robots. (2019).

[123] Giota Stratou and Louis-Philippe Morency. 2017. MultiSense—Context-aware nonverbal behavior analysis framework: A psychological distress use case. *IEEE Transactions on Affective Computing* 8, 2 (2017), 190–203.

[124] Lucy A Suchman. 2007. *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.

[125] Veikko Surakka and Jari K Hietanen. 1998. Facial and emotional reactions to Duchenne and non-Duchenne smiles. *International journal of psychophysiology* 29, 1 (1998), 23–33.

[126] Tarik Taleb, Dario Bottazzi, and Nidal Nasser. 2010. A novel middleware solution to improve ubiquitous healthcare systems aided by affective information. *IEEE transactions on information technology in biomedicine* 14, 2 (2010), 335–349.

[127] Jaime Teevan, Brent Hecht, Sonia Jaffe, Nancy Baym, Rachel Bergmann, Matt Brodsky, Bill Buxton, Jenna Butler, Adam Coleman, Mary Czerwinski, Brian Houck, Ginger Hudson, Shamsi Iqbal, Chandra Maddila, Kate Nowak, Emily Peloquin, Ricardo Reyna Fernandez, Sean Rintel, Abigail Sellen, Tiffany Smith, Margaret-Anne Storey, Siddharth Suri, Hana Wolf, and Longqi Yang. 2021. *The New Future of Work: Research from Microsoft into the Pandemic's Impact on Work Practices*. Technical Report MSR-TR-2021-1. Microsoft.

[128] Julian Thayer and Bjørn Helge Johnsen. 2000. Sex differences in judgement of facial affect: A multivariate analysis of recognition errors. *Scandinavian journal of psychology* 41, 3 (2000), 243–246.

[129] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence* 23, 2 (2001), 97–115.

[130] Marko Tkalcic, Andrej Kosir, and Jurij Tasic. 2011. Affective recommender systems: the role of emotions in recommender systems. CEUR-WS. org.

[131] Jessica L Tracy and Daniel Randles. 2011. Four models of basic emotions: a review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion review* 3, 4 (2011), 397–405.

[132] Michel F Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F Cohn. 2006. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces*. 162–170.

[133] Joelle Vanhamme and Dirk Snelders. 2001. The role of surprise in satisfaction judgments. *Journal of Consumer Satisfaction Dissatisfaction and Complaining Behavior* 14 (2001), 27–45.

[134] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[135] Stephen Wood, Valerio Ghezzi, Claudio Barbaranelli, Cristina Di Tecco, Roberta Fida, Farnese Maria Luisa, Matteo Ronchetti, and Sergio Iavicoli. 2019. Assessing the Risk of Stress in Organizations: Getting the Measure of Organizational-Level Stressors. *Frontiers in Psychology* 10 (12 2019), 2776. https://doi.org/10.3389/fpsyg.2019.02776

[136] Thomas A Wright and Barry M Staw. 1999. Affect and favorable work outcomes: two longitudinal tests of the happy–productive worker thesis. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 20, 1 (1999), 1–23.

[137] Chih-Hung Wu, Yueh-Min Huang, and Jan-Pan Hwang. 2016. Review of affective computing in education/learning: Trends and challenges. *British Journal of Educational Technology* 47, 6 (2016), 1304–1323.

[138] Tong Xue, Abdallah El Ali, Gangyi Ding, and Pablo Cesar. 2021. Investigating the Relationship between Momentary Emotion Self-reports and Head and Eye Movements in HMD-based 360 VR Video Watching. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–8.

[139] Tong Xue, Surjya Ghosh, Gangyi Ding, Abdallah El Ali, and Pablo Cesar. 2020. Designing real-time, continuous emotion annotation techniques for 360 VR videos. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[140] Jing Zhai and ARMANDO Barreto. 2008. Stress detection in computer users through non-invasive monitoring of physiological signals. *Blood* 5, 0 (2008).

[141] Biqiao Zhang and Emily Mower Provost. 2019. Automatic recognition of self-reported and perceived emotions. In *Multimodal Behavior Analysis in the Wild*. Elsevier, 443–470.

[142] Tianyi Zhang, Abdallah El Ali, Chen Wang, Alan Hanjalic, and Pablo Cesar. 2020. RCEA: Real-time, Continuous Emotion Annotation for Collecting Precise Mobile Video Ground Truth Labels. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[143] Annuska Zolyomi, Andrew Begel, Jennifer Frances Waldern, John Tang, Michael Barnett, Edward Cutrell, Daniel McDuff, Sean Andrist, and Meredith Ringel Morris. 2019. Managing Stress: The Needs of Autistic Adults in Video Calling. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–29.

[144] Shoshana Zuboff. 2015. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of information technology* 30, 1 (2015), 75–89.