emotion review

Emotion Review
Vol. 10, No. 2 (April 2018) 174–183

ISSN 1754-0739
DOI: 10.1177/1754073917696583
journals.sagepub.com/home/er

The Affective Computing Approach to Affect Measurement

Sidney D'Mello
Department of Computer Science, University of Notre Dame, USA
Department of Psychology, University of Notre Dame, USA

Arvid Kappas
Department of Psychology, Jacobs University, Germany

Jonathan Gratch
Institute of Creative Technologies, University of Southern California, USA
Computer Science Department, University of Southern California, USA

Abstract

Affective computing (AC) adopts a computational approach to study affect. We highlight the AC approach towards automated affect measures that jointly model machine-readable physiological/behavioral signals with affect estimates as reported by humans or experimentally elicited. We describe the conceptual and computational foundations of the approach followed by two case studies: one on discrimination between genuine and faked expressions of pain in the lab, and the second on measuring nonbasic affect in the wild. We discuss applications of the measures, analyze measurement accuracy and generalizability, and highlight advances afforded by computational tipping points, such as big data, wearable sensing, crowdsourcing, and deep learning. We conclude by advocating for increasing synergies between AC and affective science and offer suggestions toward that direction.

Keywords

affect detection, affective measurement, multimodal sensing, supervised classification

In an iconic scene from the classic science fiction film *Blade Runner*, the protagonist Deckard is tasked with detecting and retiring (killing) bioengineered superhumans called replicants. He successfully distinguishes replicants from humans based on their empathic responses to emotion-inducing questions, relying on a sophisticated lie-detector-like device called a Voight-Kampff machine. Wouldn't it be something if such a device was available for psychological research today? Not to detect replicants, of course, but to measure psychological states—an immense challenge given their conceptual rather than physical status. The measurement challenges are amplified in affective science because the central constructs (moods, emotions, affective states) are diffuse and have resisted clear definition and categorization for over a century (Izard, 2010; Kappas, 2013).

It may be surprising to some that the emerging field of affective computing (AC) has been working on machines that automatically measure affective states for two decades. Affective computing (Picard, 1997), broadly defined as computing involving or arising from human emotion, is an interdisciplinary field that integrates the affective and computational sciences. AC is a young field, whose origin can be traced to Picard's foundational book that named the field (Picard, 1997). The first biannual Affective Computing and Intelligent Interaction conference was held in 2005, its professional organization, the Association for the Advancement of Affective

Author note: The first author was supported by the National Science Foundation (NSF; DRL 1108845 and IIS 1523091). Any opinions, findings and conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the funding agencies.
Corresponding author: Sidney D'Mello, Departments of Computer Science & Psychology, University of Notre Dame, 118 Haggar, Notre Dame, IN 46556, USA. *Email*: sidney.dmello@gmail.com

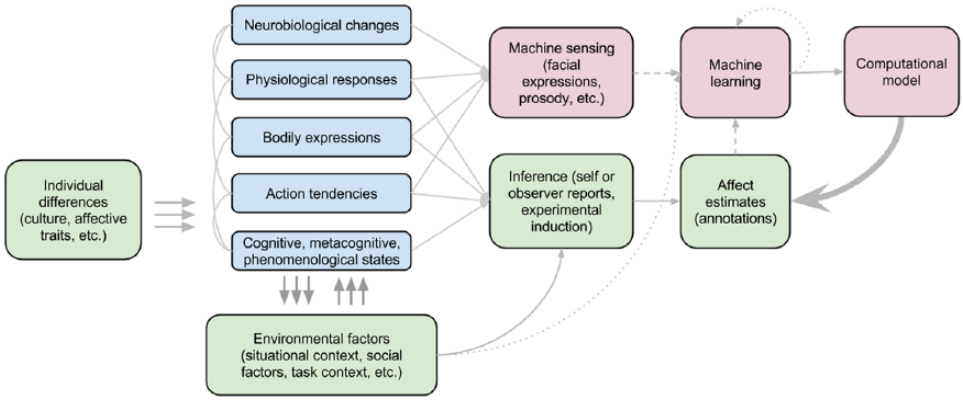

Figure 1. Conceptual and computational foundation of the AC approach.

Computing (AAAC), was founded in 2007, and its flagship journal, *IEEE Transactions on Affective Computing*, published its inaugural issue in 2010. AC has since grown into a burgeoning research area with multiple subfields, each emphasizing one or more aspects of computational emotions as discussed in edited volumes (Calvo, D'Mello, Gratch, & Kappas, 2015; Scherer, Bänziger, & Roesch, 2010).

AC owes much to affective science, where over a century of research on emotion and its influence on cognition, action, and social interaction has triggered an unprecedented interest in emotion among computational scientists and engineers. It is particularly telling that the Institute of Electrical and Electronics Engineers, the world's largest professional association, hosts the flagship journal of the field alongside more traditional publications like *Antennas and Propagation, Mechatronics*, and *Photonics*. In turn, AC can contribute to affective science due to its emphasis on advanced sensing, computational modeling, objective scalable measurement, and real-world applications (Kappas, 2010).

In this article, we focus on the AC approach to affect measurement (also called affect detection or affect recognition) to illustrate one AC research area and as a means to facilitate cross-talk with affective science. We propose that the AC measurement approach (henceforth AC approach) can offer a complementary perspective to the traditional methods used in affective science (Coan & Allen, 2007). Though they may not yet live up to the expectations of science fiction, AC measures are impressive in their own right. They go beyond merely objectively reading out bodily signals—they make inferences on the likelihoods of affective states from the signals. They can provide reliable, continual assessments of affect at fine-grained temporal resolutions with no human involvement, which potentially makes them useful to measure affect over extended periods of time, both in and outside of the lab.

In what follows, we provide an overview of the theoretical foundation of the AC approach, provide case studies, and discuss applications, strengths, weaknesses, and advances. In the interest of maximizing readership, we do not delve into mathematical formalisms or specific software tools, focusing instead at the conceptual level with illustrative examples.

Conceptual and Computational Foundation

The AC approach is conceptually grounded in affective science and in the study of psychophysiology and nonverbal behavior. Its computational footing lies in digital signal processing (Vinciarelli, Pantic, & Bourlard, 2009) and machine learning (Domingos, 2012). Figure 1 illustrates how we envision the conceptual foundation of the AC approach. Our perspective aligns with the current thinking that affective states are dynamically constructed from internal processes in a manner that is coupled with the ongoing person–environment interaction context (Kappas, 2013; Lewis, 2005; Mesquita & Boiger, 2014; also see Barrett, 2014). Further, we maintain that affective states are multicomponential, encompassing neurobiological changes, physiological responses, bodily expressions, action tendencies, and cognitive, metacognitive, and phenomenological states. These changes are modulated by contextual and social influences (Kappas, 2013; Mesquita & Boiger, 2014) and individual differences, such as affective traits and culture (Elfenbein & Ambady, 2002).

There is no box or label titled *emotion* in Figure 1. This is mindful of the current controversy on what is an emotion, a term that has been notoriously difficult to define (e.g., Izard, 2010; Kappas, 2013). Instead, the figure lists the components that most researchers would agree as being relevant for emotion without getting into whether the subjective experience is the emotion, or the whole complex. We refer to this abstract representation as the *affect estimate* (methodologically as the affect annotation), whose operationalization depends on the source, be it the self, external observers, or experimental methods. Specifically, the self has conscious access to subjective feelings, overt actions, memories of the experience, metacognitive reflections, and some physiological changes, but not to unconscious components (e.g., neurobiological changes). In contrast, external observers only have access to visible behaviors and actions (e.g., facial expressions, gestures, actions) and must rely more heavily on inference (Mehu & Scherer, 2012), including situational context (Kappas, Hess, & Scherer, 1991). An emotional response can also be experimentally elicited (Coan & Allen, 2007), but there is no guarantee that the same stimulus will

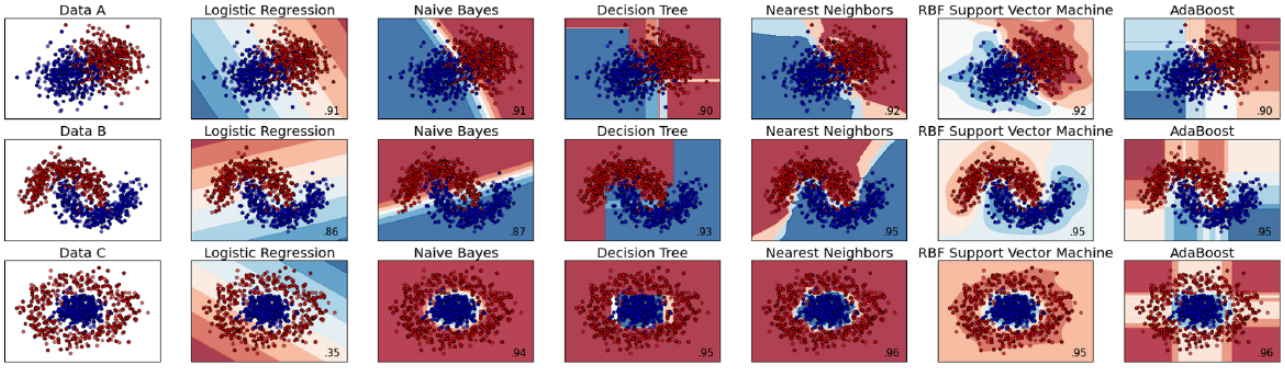

Figure 2. Decision boundaries (shaded colors) and classification accuracies (bottom right corner in each subplot) obtained by applying various supervised learning methods on three simulated data sets.
Note. RBF = radial bias function kernel.

evoke a similar response across all individuals. Thus, affect measurement requires inference, or construction, to produce an affect estimate, such as the amount of felt anger for self-reports, judged anger for observer reports, or elicited anger when experimentally induced.

Machine sensors can measure bodily/physiological signals (e.g., infrared, EEG), extending beyond what can be easily perceived by humans (e.g., facial expressions). However, they cannot infer an affect estimate from the measurements themselves. In contrast, AC measures go beyond passive sensing by *inferring* affect estimates from machine-readable signals. To do this, they assume a link between an affect estimate (provided by the self, observers, or experimentally elicited) and machine-readable bodily/physiological signals. The link need not be particularly strong, nor does there need to be strong coherence or synchrony among the various bodily/physiological responses. It is not even necessary that the link be consistent across individuals, situations, and cultures. The only assumption is that the link is "beyond-chance probabilistic" (Roseman, 2011, p. 440), reflecting a "ground truth" between machine-readable signals and affect estimates from self/observer/elicitation.

The goal of the AC approach is to computationally model this link, which requires solving two main challenges. The first is to obtain abstractions (called descriptors or features) from raw signals recorded by sensors. For example, if the signal is a facial video recorded from a camera (the sensor), the descriptors might be the activation of facial action units (AUs; Ekman & Friesen, 1978) or facial textures. Computer vision-based techniques are needed to automatically compute facial descriptors from video. Similarly, pitch and energy are common paralinguistic descriptors computed by applying acoustic processing methods to audio signals recorded with microphones. In general, automatically computing descriptors from signals falls under the purview of digital signal processing, and its subfields of computer vision, acoustic processing, computational psychophysiology, and so on. It also entails correction for measurement errors, interpolation techniques for missing data, and methods for signal synchronization, filtering, and denoising. A full description of the techniques involved is beyond the scope of this article, but we refer the reader to reviews in Calvo et al. (2015), Vinciarelli et al. (2009), and Zeng, Pantic, Roisman, and Huang (2009).

The second main challenge is to produce affect estimates from the descriptors. The most common approach uses techniques from a subfield of machine learning called supervised learning (Domingos, 2012, p. 78). Supervised learning uses *training* data consisting of descriptors that are temporally synchronized with researcher-provided affect annotations (from self-reports, observer judgments, or elicited condition) to *model* (learn) the relationship between the two (dashed lines in Figure 1). Ideally, the models should also include contextual information, which can be broadly divided into: (a) environmental factors, such as situational aspects, task constraints, and the social environment, and (b) internal context, which could include the model's own representations of affective processes and its earlier affect estimates (dotted lines). The resultant *computational model* yields computer-generated affect annotations when presented with a new set of descriptors and environmental factors without researcher-provided annotations (the thick line).

Note that we use the term computational model to specifically refer to an output of a supervised classifier; this term has broader uses in affective computing, for example to model affective processes as reviewed in Marsella, Gratch, and Petta (2010). Based on the supervised learning method, the computational model can take on many forms, such as an equation, a set of rules, a decision tree, a forest of decision trees, and a neural network. To get an intuitive sense of what these models do, and how they differ, consider the three simulated data sets shown in the left column of Figure 2. Our goal is to discriminate between two affective states (represented by 400 red and 400 blue points) from two descriptors (d1 and d2 on the axes).

We trained various standard classifiers on each data set after selecting a random 60% of the points for training and reserving the remaining 40% for testing. More specifically, after training,

the resultant model was presented with a point from the test set (with the color withheld) and its prediction of whether it was a red or blue point was compared with the actual color. We measured accuracy by computing the percent of agreement (called recognition rate or RR) between the predicted and actual color for all test set points. Random guessing (chance) would yield a RR of 50% as there are equal numbers of red and blue points.

Logistic regression, a widely used technique in the behavioral sciences (Cohen, Cohen, West, & Aiken, 2003), was extremely effective for Dataset A (RR of 91%), but RR dropped to 86% for Dataset B, and was below chance (RR of 35%) for Dataset C (see Figure 2, column 2). However, the other five models (see Figure 2, columns 3 to 7) were effective for all three data sets (see Domingos, 2012, for details on the other models). For example, a naïve Bayes classifier that uses basic Bayesian inference (Friedman, Geiger, & Goldszmidt, 1997) yielded RRs of 91%, 87%, and 94% for Datasets A, B, and C, respectively (column 3). Similarly, decision trees that utilize information theoretic heuristics (Quinlan, 1993) achieved RRs of 90%, 93%, and 95% for Datasets A, B, and C, respectively (column 4).

The performance differences between logistic regression and the other models can be attributed to how they "carve up" the data (denoted as red/blue background shading). Logistic regression uses linear decision boundaries (i.e., a line to separate the red and blue points)—an appropriate choice when the data are linearly separable (Dataset A), but a poor choice when it is not (e.g., Dataset C). In contrast, the other methods are more apt at adjusting their decision boundaries based on the data at hand, which is reflected in high accuracies across the three data sets.

Turning to the question of how to evaluate the models, we consider accuracy and generalizability as two basic criteria. There are additional application-specific evaluation criteria that are not discussed further, such as the ability to handle missing/noisy data, the ability to run in real-time, and whether the model offers inspectable representations versus being a "black box."

Accuracy (similar to convergent validity) refers to the extent to which computer-generated affect estimates align with some external standard, such as self or observer annotations or the target emotion when experimentally elicited. The simplest accuracy metric, recognition rate (RR) suffices when the data are balanced (e.g., equal red and blue points in the previous example), but is severely limited when the data are imbalanced. For example, if the task is to discriminate between anger and neutral, and 80% of the cases are neutral, a model that simply predicts neutral 100% of the time will yield a RR of 80%. Hence, researchers eschew RR in favor of alternate accuracy metrics, such as the area under the receiver operating characteristic curve (AU ROC or AUC or A-prime), Cohen's kappa, or F1 (harmonic mean of precision and recall). While these metrics are robust to some data imbalance, they are limited under conditions of extreme imbalance (Jeni, Cohn, & De La Torre, 2013). Alternate metrics, such as area under the precision-recall curve (AUPRC; Davis & Goadrich, 2006), have been proposed for these cases, but uptake has been slow. Thus, it is best to report raw classification tables, so researchers can ascertain how the reported metrics are affected by data imbalance and can calculate alternate metrics as needed.

Generalizability (or external validity) is concerned with the robustness of the model when applied to *new* data—or data different from the training data. A simple test of generalizability to new individuals involves dividing the participants in the data set into two groups, building the model on one group (training set), and testing it on the other group (testing set). Cross-validation is a widely used variant of this procedure where each group serves as training and testing data across multiple *folds* (i.e., for three folds, the data is divided into three sets X, Y, and Z. Fold 1: Train X and Y, Test Z; Fold 2: Train X and Z, Test Y; Fold 3: Train Y and Z, Test X). Additional tests of generalizability involve dividing the data by some characteristic (e.g., sex, ethnicity), training on one subset (e.g., males) and testing on the held-out subset (e.g., females). When models have internal parameters that need tuning (called hyperparameters) or when multiple models need to be compared (called model selection), the data can be divided into three subsets: training, validation, and testing, with the validation set serving as an intermediate test set for parameter tuning and model selection, so that the actual test set is uncompromised.

Case Studies

We selected two different case studies to illustrate the AC approach. The first shows how machines can outperform humans in distinguishing between genuine and faked expressions of pain in controlled lab conditions (Bartlett, Littlewort, Frank, & Lee, 2014). The second exemplifies multimodal detection of nonbasic affective states (e.g., confusion, boredom) in the wild, specifically in the noisy real-world context of a school computer lab (Bosch, D'Mello, Baker, Ocumpaugh, & Shute, 2016). These studies represent a very small sample of the research in this area (we conservatively estimate around 1,000 published studies at the time of writing), so we acknowledge that our choice of case studies is both subjective and incomplete, and direct the reader to published reviews (Calvo & D'Mello, 2010; D'Mello & Kory, 2015; Gunes & Pantic, 2010; Zeng et al., 2009).

Study 1: Discriminating Between Genuine and Faked Expressions of Pain in the Lab

Humans have considerable difficulty in discriminating real from faked expressions of some feelings, especially pain (Hadjistavropoulos, Craig, Hadjistavropoulos, & Poole, 1996). Can machines do better? Bartlett et al. (2014) addressed this question by developing an AC measure to discriminate between genuine and faked expressions of pain and compared its accuracy to humans (see Figure 3A for an overview). First, they collected 1-minute videos from 25 *stimulus subjects* during a cold-pressor pain induction task (Hadjistavropoulos et al., 1996). Genuine pain was induced by asking the stimulus subjects to submerge their arms in ice water (5 ^{o}C) for 1 minute. The same subjects were also asked to fake painful expressions while submerging their arms in warm water (20 ^{o}C). In AC parlance, the assigned experimental condition (genuine vs. fake pain) served as the affect annotation.

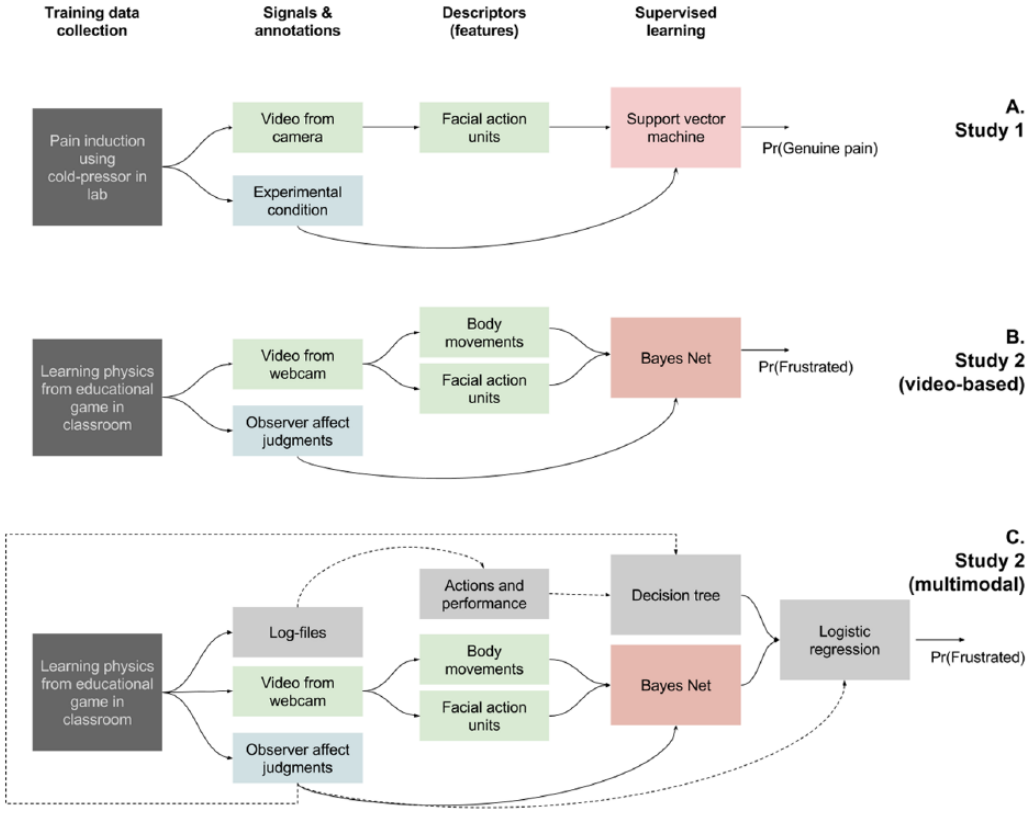

Figure 3. High-level overview of case studies.
Note. Pr. = model prediction.

The researchers used the Computer Expression Recognition Toolbox (CERT), which was trained on an independent data set (Littlewort et al., 2011), to extract time series of frame-level facial action unit (AU) activations from the videos. Next, they computed descriptors that modeled the temporal dynamics of each AU across the 1-minute video sequence. A support vector machine (Cortes & Vapnik, 1995), which is a common supervised classifier, was trained to discriminate between genuine versus faked pain expressions from the temporal descriptors. The model was cross-validated at the stimulus subject level, accomplished by training on data from all but one subject, testing on the held-out subject, and repeating until every subject was held-out once (called leave-one-subject-out cross-validation).

The researchers used the area under the ROC curve (AUC) as the accuracy metric. The ROC curve is obtained by plotting true-positives versus false-positives at various decision thresholds ranging from 100% true positives to 100% false positives (Hanley & McNeil, 1982). AUCs range from 0.5 (chance model) to 1.0 (perfect discrimination) and their model achieved an AUC of 0.91, which was statistically and substantially greater than chance.

The researchers also assessed the accuracy of human judges on the same task. In one experiment, 170 judges watched videos of the stimulus subjects in random order and judged if the expressions of pain were genuine or fake. The resultant RR of 52% was statistically indistinguishable from chance (50%). In a separate experiment, 35 different judges were first shown 24 videos of the stimulus subjects along with experimental condition (i.e., genuine vs. fake), thereby mirroring the model training procedure. When tested on 20 videos from a different set of stimulus subjects, they obtained an above-chance RR of 55%. In contrast, the computer model's RR was estimated at 85% after setting the decision threshold of the ROC curve to yield equivalent false positives and false negative rates. Thus, training helped the human judges cross the above-chance threshold, but they were still statistically and substantially less accurate than the machine.

How did the computer do it? It turned out that the pertinent information was contained in the temporal dynamics of the AUs since randomly shuffling the video frames in order to break the temporal dependencies significantly reduced the RR from 85% to 56%. Further, ignoring the temporal dynamics by focusing on the degree of activation of each AU yielded a RR of 66%, which was above chance, but lower than the 85% RR obtained when temporal dynamics were modeled. It turned out that AU 26 (mouth opening) was the most informative descriptor. Across participants, the duration and variance of mouth openings as well as the interval between consecutive mouth openings was lower for faked versus genuine pain expressions. Importantly

there was no difference in mean activation of AU 26, indicating that the computer's ability to contrast the dynamics of mouth opening across expression types was critical for its success.

Study 2: Detecting Naturalistic Episodes of Nonbasic Affect in the Wild

The second study (Bosch et al., 2016) focused on affect measurement in a noisy real-world setting of a computer-enabled classroom. The affective states of interest were boredom, confusion, delight, engagement, and frustration—these states are found to be more frequent during learning than the "basic" emotions (D'Mello, 2013) and are presumably more challenging to model because their expressive correlates have yet to be fully mapped out.

The researchers collected training data while 137 middle and high school students played a conceptual physics educational game during their assigned class periods. Game play lasted for approximately 2 hours and was spread across 2 days separated by a 3-day interval. Trained researchers performed live affect annotations during game play by observing one student at a time until visible affect was detected or 20 seconds had elapsed before moving on to the next student in a preplanned order (see Ocumpaugh, Baker, & Rodrigo, 2012). The annotations were based on body movements, gestures, facial expressions, explicit actions towards the interface, and interactions with peers and teachers. The observers had to achieve a minimum kappa of 0.6 with an "expert" prior to making the observations. Videos of students' faces and upper bodies were recorded during game play and synchronized with the affect annotations using Internet time servers.

The videos were processed using the FACET computer-vision program Version 2.1, which is a commercial version of the CERT system used in the Bartlett et al. (2014) study reviewed before. FACET provides frame-based estimates of the likelihood of 19 facial AUs along with head pose and orientation using models trained on independent data sets. The descriptors consisted of the median, standard deviation, and maximum activation of each AU, head pose, and orientation across short windows (3–12 seconds) preceding each affect annotation. Descriptors reflecting gross body movement were also computed by applying motion filtering algorithms on the videos (Kory, D'Mello, & Olney, 2015).

The researchers trained supervised classification models to discriminate each affective state from all the others (e.g., frustrated vs. other), resulting in five models, one for each state (see Figure 3B for the frustration model). The models were validated across 150 iterations of student-independent validation in which a random subset of approximately 66% students was selected for training and the remaining subset was used for testing. AUCs for the computer models exceeded the chance-level AUC of .50 for all the affective states: bored (.61), confused (.65), delighted (.87), engaged (.68), and frustrated (.63). Follow-up validation analyses confirmed that the models generalized across multiple days (i.e., training on a subset of students from Day 1 and testing on different students from Day 2; vice versa), class periods (i.e., training on a random five of the seven class periods, testing on the remaining two; repeating across multiple iterations), gender (i.e., training on males, testing on females; vice versa), and perceived ethnicity (i.e., coded by humans as no demographics were available).

Video-based measures can only be used when the face is detected in the video. This is not always the case outside of the lab, where there is little control over movement, occlusions, poor lighting, and other complicating factors. In fact, the face could only be detected about 65% of the time in this study. To address this, Bosch, Chen, Baker, Shute, and D'Mello (2015) developed an additional computational model based on descriptors extracted from information on the ongoing task context (stored in log files), such as the difficulty of the current game level attempted, the student's actions, the feedback received, response times, and so on. Then, logistic regression models were trained to adjudicate between the affect estimates of the video and interaction models, essentially weighting their relative influence on the final outcome (see Figure 3C). The resulting multimodal models were almost as accurate as the video-based models (less than 5% difference in AUC), but could be applied 98% of the time (compared to 65% for video-based models). These results are notable given the noisy nature of the real-world environment with students incessantly fidgeting, talking with one another, asking questions, and even occasionally using their cellphones.

Applications of the Models

The computational models can be used in multiple ways, beginning with measurement. Here is one use case. A researcher has collected hour-long facial videos from 500 participants and needs to code them for affective responses. However, due to the labor intensive nature of video coding, she only codes 10 randomly selected 30-second clips per video. This amounts to about one sample every 6 minutes, which might provide an indication of the affective responses in aggregate, but is too coarse-grained for an analysis of affect chronometry (i.e., time course of affective responses). To address this, she trains a computational model using automatically computed facial expressions (from videos) as descriptors and the sparse affect codes as annotations. She then applies the model on the same videos, but at a much finer temporal resolution (e.g., in 10-second intervals or 600 measurements per video compared to 10 by humans). A year later, she obtains videos from 10,000 new participants under similar experimental conditions and uses the same model to automatically generate affect annotations in a matter of hours and at no additional annotation cost.

A second application is analytic. Similar to examining the coefficients of a regression model, one can "look inside" the computational models to gain insight into how they operate. Some models are more inspectable than others. These include models that learn rules (e.g., if furrowed brow is detected then increase the likelihood of sadness), organize the rules into decision trees, or compute conditional probability tables (e.g., probability [self-report of sadness | furrowed brow and low arousal]).

Others (e.g., neural networks) are less inspectable, but some insights can be gleaned by examining individual descriptors as illustrated in the previous case study on pain detection. Additionally, an analysis of cases where the model systematically fails versus succeeds can help establish boundary conditions or identify moderating factors (e.g., Girard, Cohn, Jeni, Sayette, & De la Torre, 2015).

The AC approach can be used in simulated experiments to address questions that preclude traditional experiments. Here is an example pertaining to the debate on the universality versus (cultural) specificity in the expression and perception of basic emotions (e.g., Scherer, Clark-Polner, & Mortillaro, 2011). One could train culture-specific models on facial data collected using similar elicitation protocols but from different cultures. An analysis of the models could reveal whether a particular AU is diagnostic of either a self-report of an emotional state or the reaction to an emotionally elicited stimulus in one culture but not the other. The culture-specific models could also be compared to a universal model trained by pooling data from both cultures. Further, cross-cultural models, obtained by pairing facial videos reflecting emotional expressions from one culture with annotations from judges of a separate culture, can provide insights into whether observers from different cultures are sensitive to different facial features.

The fourth use is more application-oriented, achieved by embedding the models in real-time closed-loop systems. For example, the models can be integrated in an emotion elicitation system that customizes stimuli for individual participants based on a real-time assessment of their emotional states. They can also be integrated in intelligent interfaces that aim to increase the communicative bandwidth between the human and machine by incorporating affect into their response. Examples of these affect-sensitive or affect-aware intelligent systems include educational systems that dynamically tailor their instructional activities based on student affect (D'Mello & Graesser, 2015), entertainment systems that automatically customize stimuli to induce specific moods (van der Zwaag, Janssen, & Westerink, 2012), and advertising/marketing applications that customize advertisements based on viewer affect (McDuff, Kaliouby, Cohn, & Picard, 2015).

Analysis of the Approach

How do the AC measures perform with respect to the criteria of accuracy and generalizability? To preview our analysis, AC measures are not "magical" devices that can perfectly "read-out" an affective state from measurable signals for everyone and in every situation—and this is not the goal as we see it. In our view, it is more productive to establish meaningful lower and upper bounds rather than obsess over absolute criteria for accuracy and generalizability.

In terms of accuracy, the current standard is to show that a model is *above-chance* accurate. This is a reasonable starting point given the weak relationships between measurable bodily signals, affective states, and loose coupling between different components of an affective response (as discussed before). Further, supervised classification requires affect annotations from the self, observers, or the elicited condition. The annotation process is imperfect, thereby introducing additional variance to the mix. Further, if the phenomenon (i.e., links between affect elicitation, expression, experience, and perception) itself is only "beyond-chance probabilistic" (Roseman, 2011, p. 440), then so will be attempts to computationally model it.

Defining an upper bound is less clear-cut, because accuracy depends on the complexity of the affective states being modeled and on the variability and size of the training data. In particular, more basic feeling states (e.g., pain, pleasure) might be easier to model than states like interest, curiosity, frustration, and pride. Accuracy should also be higher when there is less variability in the data, for example, when emotions are experimentally induced in the lab compared to when they naturally occur in the wild. Greater variability requires a proportional increase in the amount of training data required, so data availability also comes into play. Thus, although it is difficult to firmly establish an accuracy upper bound, near-perfect accuracy ought to be viewed with some skepticism because either the phenomena had been considerably diluted (e.g., asking actors to pose facial expressions depicting particular emotions) or the model has been "overfit" to the data (e.g., improper validation may have led to homogenous training and testing sets) and is unlikely to generalize.

Generalizability has been a moving target. In the early days, it was sufficient to demonstrate that a model would generalize to new data, usually from the same person. In other words, training and testing data sets consisted of different cases, but the same person could be represented in both data sets, which prevents making claims of generalizability beyond the individual. The current standard is to perform *person-independent* validation, where data from a given individual can be in either the training or testing set, but not both. This is quite a challenge due to individual differences in affective responses and in some bodily signals (notably physiology).

An essential next step is to provide evidence for population generalizability (Ocumpaugh, Baker, Gowda, Heffernan, & Heffernan, 2014), possibly by showing that the models generalize across different subsets of the population, including appearance (e.g., facial hair or not; wearing glasses or not), gender, age, ethnicity, and culture. It might also be necessary to build individual models for each subset if a universal model fails to adequately generalize.

Generalizability across time, or temporal generalizability, of the measures also needs more attention. We know that baseline physiological responses can vary considerably from day to day, so generalization to data from the same individual but across different days can be challenging (Picard, Vyzas, & Healey, 2001). Bosch et al. (2016) showed that their model trained on one set of individuals generalized to data from different individuals collected 3 days later, but what about generalizability across weeks, months, and even years?

Situational generalizability is perhaps the most understudied aspect of AC measures. The current approach is to collect training data from one situational context with specific affordances,

such as affective-inducing events, affective states, affective expressions, instrumental actions, and social interactions (or lack thereof). It is quite possible that the models might be overfit to the training context and are unlikely to generalize to new contexts. For example, a model trained to measure anger while viewing anger-eliciting films might degrade when applied to different contexts, such as a road-rage driving scenario or during an argument with a coworker. Similarly, changes in affective expressions as a function of the social environment (e.g., presence of a superior) would likely cause measurement error because the AC measures do not (yet) model social context in any meaningful way.

Advances to the Approach

The first-generation AC measures (*c*.2000) served as a proof-of-concept in that they mainly focused on unimodal measurement of posed expressions of basic emotions in very restricted contexts with little concern for generalizability. We are currently in the second generation of measurement, where the emphasis is on multimodal modeling of a more diverse set of affective states in less restrictive contexts and with the expectation of person-level generalizability. We anticipate the third generation of AC measures will be characterized by major advances afforded by various technological tipping points as we elaborate next.

The era of "big data" is one tipping point. An oft-learned lesson of machine learning is that, all else being equal, more data beats smarter algorithms (Domingos, 2012). Accordingly, the current practice of collecting data from small samples (usually < 100 individuals) who engage in one or two carefully crafted activities is insufficient to develop highly accurate or generalizable models. But it need not be this way anymore. The ubiquity of webcams and microphones in everyday devices, the availability of wearable sensing devices (e.g., wrist bands for multichannel physiology), cost-effective alternatives to research-grade sensing (e.g., Fitbit for physiology, EyeTribe for eye gaze, and Kinect for motion capture), the ability to sense environmental and social context (e.g., smartphones with GPS), and cost-effective computation and storage on the cloud affords the collection of unprecedented amounts of naturalistic affective data from diverse individuals on the go, across a range of situations, day and night, every day and every night. Thus, one advance involves models that are trained on a wide range of individuals' emotional experiences as they go about their daily lives (Picard, 2010).

There are clearly challenges that need to be addressed along the way. At first blush, the cost of some physiological sensors is a barrier to scalability; however, some of these sensors can be replaced with scalable proxies. Cameras are the ideal proxy sensor as they are integrated in most 21st-century generation computing devices and can be used to track a range of signals, such as facial activity (e.g., Littlewort et al., 2011), bodily movements (Kory et al., 2015), heart rate (Poh, McDuff, & Picard, 2010), and even eye gaze (Sewell & Komogortsev, 2010). Further, variability in the color images of a fingertip can be used to estimate several physiological signals including respiration rate, heart rate variability, and blood oxygen saturation (e.g., Scully et al., 2012). When combined with online data collection platforms like Amazon's Mechanical Turk (MTurk), proxy sensing affords unprecedented data collection from individuals situated all over the world. Of course, these sensing methods are often ill-suited to assess the ongoing social context (who is present, what is the relationship between these people), but this can be partly remedied with advances in context modeling (Bettini et al., 2010). Further, privacy concerns that emerge when people's homes become part of the measurements need to be addressed with appropriate safeguards and by seeking informed consent (see Cowie, 2015, for a discussion of ethics in AC).

There is still the critical challenge of annotating all of this "big" data as the laborious lab-based annotation methods do not scale. To address the annotation bottleneck, researchers have begun to leverage the power of social media and human intelligence platforms like MTurk for crowdsourced affect annotation (Morris & McDuff, 2015). Although such annotations are noisier than lab-based annotations, similar levels of *effective reliability* can be achieved by increasing the number of annotators per stimulus (Rosenthal, 2005), which is feasible as per-stimulus annotation costs are in the order of cents rather than the dollars needed for lab-based annotation. Additionally, semisupervised learning methods (Zhu, 2005), which only require a small subset of the data to be annotated, provide alternatives when it is infeasible to annotate the entire data set.

Finally, there have been tremendous advances in the machine learning methods themselves. Discriminative probabilistic graphical models, such as hidden conditional random fields, boosted-coupled hidden Markov models, and dynamic Bayesian networks (Koller & Friedman, 2009), offer impressive levels of modeling sophistication, such as hierarchical relationships (e.g., how affective traits influence affective expressions), interdependencies among descriptors (e.g., how facial expressions co-occur), and time-lagged temporal dependencies (e.g., how affect at time t influences affect at $t+1$). Complementary advances are emerging from the field of *deep learning* (Le Cun, Bengio, & Hinton, 2015), where models such as convolutional neural networks aim to learn higher level abstractions from low-level input by incorporating multiple processing layers, each performing a transformation of the representation from the previous layer. The most exciting aspect of these methods is that the abstractions (descriptors) dynamically emerge during learning rather than being precomputed, which has considerable theoretical and practical implications.

Facilitating Cross-Talk Among Disciplines

How might the fields of affective computing (AC) and affective science collaborate in the future? We hope that affective science researchers would consider adopting the AC measurement approach due to its many advantages and applications as elaborated in this article. We acknowledge that the measures are imperfect, but we consider this unsurprising as affect measurement involves inference of a complex psychological construct, an inherently imperfect endeavor irrespective of whether the

inference is done by humans or machines. In our view, it is more efficacious to leverage the most out of imperfect measures than to await perfect measurement. Indeed, some of the sources of measurement error can be remedied by increased cross-talk among the disciplines. In particular, AC has been slow to adopt the numerous advances from affective science; consequently, many AC measures are rooted in outdated affect theory and methodology (D'Mello & Kory, 2015; Kappas, 2010). AC's responsiveness to advances from its core engineering and computational fields needs to be matched with corresponding receptiveness to theoretical and methodological advances from affective science—a feat that can easily be achieved when AC and affective science researchers work hand in hand.

The easiest way for the two disciplines to collaborate in the near-term is by incorporating research methods and artifacts (i.e., tools, data) from each other. Some of this is already underway. AC has extensively used affect elicitation techniques from affective science (Kory & D'Mello, 2015), while affective science is beginning to incorporate the AC measurement approach. For example, Kragel and LaBar (2013) used support vector machines to build models that discriminated emotions from multichannel peripheral physiology, while Laukka, Neiberg, and Elfenbein (2014) also used support vector machines to study cross-cultural versus within-cultural emotion classification from acoustic descriptors (similar to one of the application areas discussed in the previous lines).

The AC community has developed several freely available artifacts that can serve as tangible resources to catalyze cross-talk among the two fields. These include the approximately 50 databases (Association for the Advancement of Affective Computing [AAAC], 2016) featuring a variety of affective phenomena collected from multiple modalities and across a range of contexts along with software tools to analyze the data (Hussain, D'Mello, & Calvo, 2015). As one example, given the importance of facial expressions to emotion research and the labor-intensive nature of manual Facial Action Coding System (FACS) coding, automated facial expression analysis software (Girard et al., 2015) has the potential to be the "killer" application to unite the two fields.

At a more basic level, joint activities, such as conferences or symposia at conferences anchored in the respective disciplines could link researchers to form interdisciplinary teams that support each other and complement the respective strengths of the other discipline(s). Similarly, interdisciplinary training of students would be a useful strategy, such as international research training networks that attempt to make the necessary connections.

Concluding Remarks

We presented an overview of the affective computing (AC) approach to affect measurement as one means to stimulate cross-talk between AC and affective science. Affective science provides the theoretical inspiration for AC, which can provide the technological triggers to radically advance affective science. Like computational neuroscience, computational finance, computational chemistry, and computational biology have enriched their respective core fields, computational affect can similarly enrich the study of emotion. However, transformative innovation will only be possible when AC and affective science researchers work together, each challenging and enriching the other's world view, methods, and tools. Indeed, the two fields can greatly benefit from a mutual symbiosis given their shared underlying fascination with emotion.

Declaration of Conflicting Interests

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

References

Association for the Advancement of Affective Computing (AAAC). (2016). Collection of emotional databases [Databases]. Retrieved from http://emotion-research.net/wiki/Databases

Barrett, L. F. (2014). The conceptual act theory: A précis. *Emotion Review*, *6*, 292–297.

Bartlett, M. S., Littlewort, G. C., Frank, M. G., & Lee, K. (2014). Automatic decoding of facial movements reveals deceptive pain expressions. *Current Biology*, *24*(7), 738–743.

Bettini, C., Brdiczka, O., Henricksen, K., Indulska, J., Nicklas, D., Ranganathan, A., & Riboni, D. (2010). A survey of context modelling and reasoning techniques. *Pervasive and Mobile Computing*, *6*(2), 161–180.

Bosch, N., Chen, H., Baker, R., Shute, V., & D'Mello, S. K. (2015). Accuracy vs. availability heuristic in multimodal affect detection in the wild. *Proceedings of the 17th ACM International Conference on Multimodal Interaction* (pp. 267–274). New York, NY: ACM.

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., & Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems*, *6*(2). doi:10.1145/2946837

Calvo, R. A., & D'Mello, S. K. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*(1), 18–37. doi:10.1109/T-AFFC.2010.1

Calvo, R., D'Mello, S. K., Gratch, J., & Kappas, A. (Eds.). (2015). *The Oxford handbook of affective computing*. New York, NY: Oxford University Press.

Coan, J., & Allen, J. (Eds.). (2007). *Handbook of emotion elicitation and assessment*. New York, NY: Oxford University Press.

Cohen, P., Cohen, J., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cowie, R. (2015). Ethical issues in affective computing. In R. Calvo, S. K. D'Mello, J. Gratch, & A. Kappas (Eds.), *The Oxford handbook of affective computing* (pp. 334–348). New York, NY: Oxford University Press.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233–240). New York, NY: ACM.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology*, *105*(4), 1082–1099.

D'Mello, S., & Graesser, A. (2015). Feeling, thinking, and computing with affect-aware learning technologies. In R. Calvo, S. D'Mello, J. Gratch, & A. Kappas (Eds.), *The Oxford handbook of affective computing* (pp. 419–434). New York, NY: Oxford University Press.

D'Mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, *47*(3), 41–46.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Ekman, P., & Friesen, W. (1978). *The Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Elfenbein, H., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*(2), 203–235. doi:10.1037//0033–2909.128.2.203

FACET – Facial Expression Recognition Software (Version 2.1) [Computer software]. Boston, MA: Emotient.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*(2–3), 131–163.

Girard, J. M., Cohn, J. F., Jeni, L. A., Sayette, M. A., & De La Torre, F. (2015). Spontaneous facial expression in unscripted social interactions can be measured automatically. *Behavior Research Methods*, *47*(4), 1136–1147.

Gunes, H., & Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, *1*(1), 68–99.

Hadjistavropoulos, H. D., Craig, K. D., Hadjistavropoulos, T., & Poole, G. D. (1996). Subjective judgments of deception in pain expression: Accuracy and errors. *Pain*, *65*(2), 251–258.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36.

Hussain, M. S., D'Mello, S. K., & Calvo, R. A. (2015). Research and development tools in affective computing. In R. Calvo, S. D'Mello, J. Gratch, & A. Kappas (Eds.), *The Oxford handbook of affective computing* (pp. 349–358). New York, NY: Oxford University Press.

Izard, C. (2010). The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review*, *2*, 363–370. doi:10.1177/1754073910374661

Jeni, L., Cohn, J., & De La Torre, F. (2013). Facing imbalanced data – Recommendations for the use of performance metrics. In A. Nijholt, S. K. D'Mello, & M. Pantic (Eds.), *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 245–251). Washington, DC: IEEE.

Kappas, A. (2010). Smile when you read this, whether you like it or not: Conceptual challenges to affect detection. *IEEE Transactions on Affective Computing*, *1*(1), 38–41.

Kappas, A. (2013). Social regulation of emotion: Messy layers. *Frontiers in Psychology*, *4*(51). doi:10.3389/fpsyg.2013.00051

Kappas, A., Hess, U., & Scherer, K. (1991). Voice and emotion. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of nonverbal behavior* (pp. 200–238). New York, NY: Cambridge University Press.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: MIT Press.

Kory, J., & D'Mello, S. K. (2015). Affect elicitation for affective computing. In R. Calvo, S. D'Mello, J. Gratch, & A. Kappas (Eds.), *The Oxford handbook of affective computing* (pp. 371–383). New York, NY: Oxford University Press.

Kory, J., D'Mello, S. K., & Olney, A. (2015). Motion tracker: Camera-based monitoring of bodily movements using motion silhouettes. *PLoS ONE*, *10*(6). doi:10.1371/journal.pone.0130293

Kragel, P. A., & LaBar, K. S. (2013). Multivariate pattern classification reveals autonomic and experiential representations of discrete emotions. *Emotion*, *13*(4), 681–690.

Laukka, P., Neiberg, D., & Elfenbein, H. A. (2014). Evidence for cultural dialects in vocal emotion expression: Acoustic classification within and across five nations. *Emotion*, *14*(3), 445–449.

Le Cun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*, 436–444.

Lewis, M. D. (2005). Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and Brain Sciences*, *28*(2), 169–245.

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). The computer expression recognition toolbox (CERT). *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition* (pp. 298–305). Washington, DC: IEEE.

Marsella, M., Gratch, J., & Petta, P. (2010). Computational models of emotion. In K. R. Scherer, T. Bänziger, & E. Roesch (Eds.), *A blueprint for a affective computing: A sourcebook and manual* (pp. 21–46). Oxford, UK: Oxford University Press.

McDuff, D., Kaliouby, R. E., Cohn, J. F., & Picard, R. (2015). Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads. *IEEE Transactions on Affective Computing*, *6*(3), 223–235.

Mehu, M., & Scherer, K. (2012). A psycho-ethological approach to social signal processing. *Cognitive Processing*, *13*(2), 397–414.

Mesquita, B., & Boiger, M. (2014). Emotions in context: A sociodynamic model of emotions. *Emotion Review*, *6*, 298–302.

Morris, R., & McDuff, D. (2015). Crowdsourcing techniques for affective computing. In R. Calvo, S. D'Mello, J. Gratch, & A. Kappas (Eds.), *The Oxford handbook of affective computing* (pp. 384–394). New York, NY: Oxford University Press.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining: A case study in affect detection. *British Journal of Educational Psychology*, *45*(3), 487–501.

Ocumpaugh, J., Baker, R. S., & Rodrigo, M. M. T. (2012). *Baker–Rodrigo Observation Method Protocol (BROMP) 1.0. Training manual Version 1.0*. Retrieved from http://www.columbia.edu/~rsb2162/BROMP%20QFO%20Training%20Manual%201.0.pdf

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Picard, R. W. (2010). Emotion research by the people, for the people. *Emotion Review*, *2*, 250–254.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175–1191.

Poh, M. Z., McDuff, D. J., & Picard, R. W. (2010). Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, *18*(10), 10762–10774.

Quinlan, J. R. (1993). *C4. 5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Roseman, I. J. (2011). Emotional behaviors, emotivational goals, emotion strategies: Multiple levels of organization integrate variable and consistent responses. *Emotion Review*, *3*, 434–443.

Rosenthal, R. (2005). Conducting judgment studies: Some methodological issues. In J. A. Harrigan, R. Rosenthal, & K. R. Scherer (Eds.), *New handbook of methods in nonverbal behavior research (Series in Affective Science)* (pp. 199–236). New York, NY: Oxford University Press.

Scherer, K. R., Bänziger, T., & Roesch, E. (Eds.). (2010). *A blueprint for affective computing: A sourcebook and manual*. New York, NY: Oxford University Press.

Scherer, K. R., Clark-Polner, E., & Mortillaro, M. (2011). In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. *International Journal of Psychology*, *46*(6), 401–435.

Scully, C. G., Lee, J., Meyer, J., Gorbach, A. M., Granquist-Fraser, D., Mendelson, Y., & Chon, K. H. (2012). Physiological parameter monitoring from optical recordings with a mobile phone. *IEEE Transactions on Biomedical Engineering*, *59*(2), 303–306.

Sewell, W., & Komogortsev, O. (2010). Real-time eye gaze tracking with an unmodified commodity webcam employing a neural network. *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems* (pp. 3739–3744). New York, NY: ACM.

Van der Zwaag, M., Janssen, J., & Westerink, J. (2012). Directing physiology and mood through music: Validation of an affective music player. *IEEE Transactions on Affective Computing*, *4*(1), 57–68.

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743–1759.

Zeng, Z., Pantic, M., Roisman, G., & Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39–58.

Zhu, X. (2005). *Semi-supervised learning literature survey*. Madison: University of Wisconsin, Madison.