The Conceptualization and Measurement of Emotion in Machine Learning: A Critical Appraisal and Recommendations from Psychology

Pablo Caceres

October 6, 2020

Abstract

While psychologists, neuroscientists, and philosophers continue to debate about the nature of human emotions, machine learning researchers hurry to develop artificial systems capable of recognizing and synthesizing (i.e., artificially generating) emotions. Such efforts have been primarily motivated by the vast space of potential applications of emotion recognition and generation systems. Applications like personalized advertising, machine-assisted education, machine-assisted psychotherapy, employee assessment, elder care robots, machine-assisted mental health diagnosis, and emotion responsive gaming, are just a couple of examples. In this context, I aim to accomplish the following objectives: (1) to review the literature on emotion from a *conceptual* perspective, this is, how the concept of emotion has been understood and operationalized in the fields of psychology and machine learning; (2) to critically examine the machine learning literature regarding the conceptualization and measurement of emotion; (3) to identify areas of improvement and innovation in the conceptualization and measurement of emotion for basic and applied research, with special attention to the needs of machine learning researchers.

Contents

1 Introduction

Emotion has been subject of speculation for philosophers and scientists for millennia (Barrett, 2017a; Sorabji, 2000). In Ancient Greece, the Stoics regarded emotion as a form of mental judgments and attitudes, particularly the mistaken ones (Sorabji, 2000). In the Middle Ages, Scholastic philosophers like Thomas Aquinas[1] considered emotions as "a movement of the sense appetite caused by imagining good or evil." (as cited in Lombardo, 2011, p. 20), whereas Enlightenment philosophers like René Descartes (Descartes, 1649) thought of *passions* (which today we may call emotions) as "perceptions, feelings, or emotions of the soul which we relate specially to it, and which are caused, maintained, and fortified by some movement of the spirits" (p., 344).

Since the advent of scientific psychology in the XIX century (Hergenhahn & Henley, 2013), psychologists and philosophers have continued debating about the nature of emotions without reaching consensus (Barrett et al., 2007; P. E. Ekman & Davidson, 1994; Griffiths, 2013; Izard, 2007; Lindquist, Siegel, Quigley, & Barrett, 2013; Scarantino, 2012), in spite of the over a hundred years of accumulated research on the subject. Today, in the dawn of the XXI century, emotion has become one the focus of attention of the machine learning (ML) and artificial intelligence (AI) communities (Calvo, D'Mello, Gratch, & Kappas, 2015; Picard, 2000). In particular, ML/AI researchers have invested significant effort developing systems for two main purposes: (1) *emotion recognition* (Cowie et al., 2001), and (2) *emotion synthesis* (i.e., artificially generated emotions) (Gunes, Schuller, Pantic, & Cowie, 2011). Such efforts have given rise to the field of *affective computing* (Calvo et al., 2015; Picard, 2000), which has been defined as "computing that relates to, arises from, or deliberately influences emotions." (Picard, 2000, p., 3)

The range of topics encompassed by the affective computing field exceeds what is typically conceptualized as *emotion* in the psychological literature. Just to mention a few examples, top-

[1]It is important to hold in mind that Aquinas, and intellectuals previous to the modern scientific era more generally, never wrote about the specific concept of "emotion". Rather, as Lombardo (2011) notice in the case of Thomas Aquinas, they wrote about *appetites*, *passions*, *affections*, and other related concepts.

ics such as: *mood* (Hernandez, Hoque, Drevo, & Picard, 2012), *sentiment* (Pang & Lee, 2008), *stress* (Healey & Picard, 2005), *wellbeing* (Nosakhare & Picard, 2020), and *pain* (Martinez, Rudovic, & Picard, 2017), all fall under the scope of affective computing research. Even considering the wide diversity of perspectives in psychology, the aforementioned topics typically are regarded as *emotion-related* or *components of emotion processes*, but not equivalent to *emotion*. This is no surprise if we consider the definition of *affective computing* given above, which makes explicit the interest on emotion-related topics in general. That being said, I believe that conflating emotion, as understood in the psychological literature, with constructs as *mood*, *stress*, *anxiety*, and others, can be detrimental to the progress and application of affective computing technology.

Consider critical applications of affective computing technology like robot-assisted therapy of children with Autism (Marinoiu, Zanfir, Olaru, & Sminchisescu, 2018; Rudovic, Lee, Mascarell-Maricic, Schuller, & Picard, 2017). In this type of setting, the *conceptualization* of emotion plays a critical role in the design and performance of therapeutic devices, and consequently, in the health outcomes of children subject to this therapeutic approach. Further, the conceptualization of emotion will outline the limitations and potentialities of affective computing technology: what is feasible and what is *not* given our current knowledge of the subject.

This article aims to accomplish the following goals: (1) to review the literature on emotion from a *conceptual* perspective, this is, how the concept of emotion has been understood and operationalized in the fields of psychology and machine learning; (2) to critically examine the machine learning literature regarding the conceptualization and measurement of emotion; (3) to identify areas of improvement and innovation in the conceptualization and measurement of emotion for basic and applied research, with special attention to the needs of machine learning researchers. Consequently, I will not examine the neuroscience literature on emotion, although references will be introduced when deemed appropriate. This is not to say the neuroscience research on emotion is not relevant for our discussion. It is. However, since our focus is *conceptual* and *descriptive*, the discussion about the neurobiological basis of emotion can be safely omitted for the most part. In other words, I am aiming to communicate how emotion

scientists have understood the *construct of emotion* and what sort of implications this may have in applied settings.

I will not attempt to be comprehensive but to focus on what my interpretation of the literature indicates as the main conceptual trends in psychology and machine learning regarding emotion. The structure of this review is as follow: first, I begin by clarifying the difference between emotion and affect; second, I describe the main theories of emotion in the contemporary psychological literature; third, I analyze the complexities of establishing the so-called *ground truth* of emotion; fourth, I review how emotion has been operationalized in the machine learning literature; fifth; I examine the main issues associated to the main trends on the conceptualization and measurement of emotion in machine learning; finally, I finish by identifying avenues for improvement and innovation in the conceptualization and measurement of emotion.

2 What emotions are not

Although there is no hard consensus about what emotions are, most authors distinguish emotions from other psychological constructs like *personality traits*, *mood*, *mental disorders*, and *general affective states*. In particular, I will focus the attention on the difference between emotion and *affect*, since these concepts are tightly related and therefore easily conflated.

As with emotion, there is not a universally accepted definition of *affect*. One of the first definitions was proposed by Wilhelm Wundt, who understood affect as a momentary mental state, characterized by feelings of *pleasantness/unpleasantness*, *rousing/subduing*, and *strain/relaxation* (Wundt, 1907). According to Wundt, these fleeting states can be considered as "basic" or "primitive" (i.e., irreducible) ingredients of the human mind. Wundt's perspective has influenced contemporary views under the rubric of *core affect* ("core" in the sense of "basic" or "primitive"). According to Russell (2003), *core affect* is "A neurophysiological state that is consciously accessible as a simple, nonreflective feeling that is an integral blend of hedonic (pleasure–displeasure) and arousal (sleepy–activated) values" (p., 147). Hence, affect is a fundamental part of emotional events, but emotions can not be reduced to affect. Consider

this quote from the poet Derek Walcott (1986):

> The time will come when, with elation, you will greet yourself arriving at your own door, in your own mirror, and each will smile at the other's welcome. (p., 328)

In this fragment, *elation* is accompanied by the *face expression* of a smile. Similarly, most if not all emotional experiences will involve other forms of behavior in addition to the fleeting states of *hedonic valence* and *arousal*. Russell and Barrett (1999) make this clearer by distinguishing between *prototypical emotional episodes* and core affect. On their view, prototypical emotional episodes include core affect, overt behavior (e.g., the smile on the mirror when arriving at your own door), attention towards, appraisal of, and attributions towards the instance triggering the emotional episode (i.e., metacognition), and the compound of neurochemical and physiological reactions underlying these psychological events.

It can be objected dimensional perspectives on emotion place hedonic valence and arousal at the center of the construct of emotion (Russell, 2003). To this, I would answer that there is a difference between *necessary* and *sufficient* conditions for the emotional experience. In other words, affective states may be regarded as *necessary* conditions of emotion, but not as *sufficient*.

The notion that theorists like Russell sustain emotions can be completely accounted for by affective dimensions, probably comes from an early study of Russell and Mehrabian (1977), where they indeed tested the hypothesis that emotion could be accounted for three orthogonal dimensions: *pleasure-displeasure*, *arousal-nonarousal*, *dominance-submissiveness*. However, Russell's theoretical and empirical work (Russell, 2003) took a shift since then towards highlighting the complexity of emotional experiences *beyond* affective dimensions as described above.

3 Theories of emotion: the emotion wars

Controversy has dominated the psychological and neuroscientific landscape of emotion theory in the last decades (Barrett et al., 2007; Cacioppo & Gardner, 1999; P. E. Ekman & Davidson,

1994; Fox, Lapate, Shackman, & Davidson, 2018; Izard, 2007, 2009; Lindquist et al., 2013). Several of the disagreement among emotion theorists are deep, as deep as the disagreements about other controversial subjects in psychology like *nature* versus *nurture* (Elman et al., 1998; Mareschal et al., 2007; Pinker, 2003a), or *distributed* versus *symbolic* cognition (McClelland, Rumelhart, & Group, 1986; Pinker, 2003b). My reading of the literature indicates that the main cleavages dividing the field are around the following issues:

- *Are emotions natural kinds?* Another way to put this is: Are emotions "discovered" or "constructed"?
- *Do emotions have predefined functions determined by our evolutionary history?* For instance: Does "anger" have a predefined evolutionary function like removing obstacles or confronting predators?
- *Do emotions have objective biologically-based markers?* For instance: Are there universally shared facial expressions that serve as objective "markers" of happiness or sadness?
- *How are emotions represented in the human mind?* For instance: Do discrete categories have enough flexibility to represent the wide range of emotional experiences across cultures?

These cleavages are not mutually exclusive and do not exhaust the points of controversy in the field (Fox et al., 2018). For instance, researchers that favor the idea of emotions as natural kinds, also favor the idea of emotions as having evolutionary predetermined functions. Yet, I hope having these questions in mind may serve as a compass to navigate the diversity of views in the literature. In our review, I identified at least four main traditions in emotion theory: (1) the *basic emotion theory*; (2) the *circumplex theory of core affect*; (3) the *functionalist theory of emotion*; (4) the *theory of constructed emotion*. In what follows, I examine these perspectives and how they relate to the above-mentioned issues. My primary goal is to describe and analyze the main ideas of each perspective, rather than recapitulating the accumulated empirical evidence in the field. Such an endeavor is beyond the scope of this review. For

broader discussion about the relative merits of each perspective on emotion see (Barrett, 2017a; J. LeDoux, 2012; P. E. Ekman & Davidson, 1994; Fox et al., 2018).

Theory	Natural Kind	Evolutionary	Objective Markers	Representation
Basic Emotion	Yes	Yes	Yes	Discrete
Circumplex	No	No	No	Continuous
Constructed emotion	No	No	No	Continuous
Functionalist emotion	Yes	Yes	Yes	Discrete Continuous

Table 1: Summary psychological theories on emotion

3.1 Basic emotion theory

In the opening scene of the American crime television drama *Lie to me**, we witness the encounter between a crime suspect and Dr. Cal Lightman, a renowned expert in the science of interpreting micro-expression and body language. In the scene, the camera carefully traces every microscopic body movement of the suspect: the quivers of his eyebrow, the pressing of his lips, the twitching of his finger, the heavy breathing of his chest. For Dr. Lightman, every micro-expression reveals a piece of the thoughts and emotions of the man in front of him, knowledge that Dr. Lightman will "weaponize" to carve out the truth about the suspect's past actions.

The science behind Dr. Lightman skills on *Lie to me**, is based on the work of Paul Ekman (P. Ekman & Friesen, 1971; P. Ekman & Keltner, 1997; P. Ekman, 1993; P. Ekman & Cordaro, 2011). Actually, not only the science, but Dr. Lightman itself is loosely based on Paul Ekman persona (P. Ekman, 2020). Of course, *Lie to me** is not a literal recreation of Ekman work, but the *core idea* behind Ekman work is what serves as inspiration: emotions as *discrete* mental states, that can be cataloged into a finite set of *classes*, and that can be recognized by stereotypical, universally shared, *physiological responses* and *facial expressions* (P. Ekman, Friesen, & Ellsworth, 2013; P. Ekman & Keltner, 1997; P. Ekman & Cordaro, 2011). From this perspective, emotions qualify as *natural kinds*: as a class of natural phenomena that exist independently of human conventions (Barrett, 2006; Izard, 2007; Scarantino, 2012). Through-

out history, different societies may have *labeled* such phenomena differently, but its existence is not subject to human arbitrariness. In this sense, emotions are no different than elementary particles in physics: as fermions and bosons, emotions are *discovered*, not constructed. Consequently, if emotions are natural kinds, there must a way to *objectively* identify them by the means of science, and indeed, methods have been developed along this line of reasoning for such purposes (P. Ekman et al., 2013). This view is what is refereed in the literature as *basic emotion theory*. Although Ekman view is probably the most prominent in this tradition, others authors like Izard (Izard, 2007, 2009), Panksepp & colleagues (Panksepp & Watt, 2011; Panksepp & Biven, 2012), and Levenson (Levenson, 1994, 2011), have also proposed variations of the basic emotion perspective.

The exact number of basic emotion classes has changed over time, but today is common to see authors referring the following (P. Ekman & Cordaro, 2011; Tracy & Randles, 2011): (1) *anger*, (2) *fear*, (3) *surprise*, (4) *sadness*, (5) *disgust*, (6) *contempt*, (7) *happiness*. The evidence supporting basic emotion views comes primarily from studies where participants attempt to infer emotional states from static images of prototypical facial configurations of emotion (Barrett, Adolphs, Marsella, Martinez, & Pollak, 2019). For instance, in their original work, Ekman and Friesen (1971) exposed 189 adults and 130 children from the South East Highlands of New Guinea to stories and photographs of faces depicting one of their hypothesized basic emotions. The participant were asked to pair stories with emotional content, with photographs of faces depicting the emotion contained in the story, according to the predefined set of: *happiness*, *anger*, *sadness*, *disgust*, *surprise*, and *fear*. Given that a similar pattern of responses was found between participants in the study and participants from previous studies in Westernized cultures, Ekman and Friesen interpreted their results as evidence of the universality of facial expression of emotion. Since Ekman and Friesen's original study, detailed descriptions of their hypothesized basic emotions have been developed, allowing researchers to identify, catalog, and simulate (with actors or computationally) emotions, and even trigger emotional experiences in laboratory experiments (Reisenzein, Studtmann, & Horstmann, 2013; Siedlecka & Denson, 2019).

It is important to remark that the basic emotion perspective theory does not propose that emotions are monolithic or inflexible. Emotions do have variation and can be shaped (to some degree) by life experiences and culture (P. Ekman & Cordaro, 2011). In this sense, basic emotions are better understood as "emotion families" with some degree of within-family variation. However, such variation is constrained by a core of evolutionary crafted basic emotions, which are shared by all humans as species (Izard, 2007). The reason why emotions have such a strong predetermined foundation according to this view is related to their hypothetical value to address threats and opportunities over evolutionary scale (Izard, 2007; Panksepp & Biven, 2012)

For an extended review of the basic emotion theory perspective, refer to the works of Izard (2007, 2009), Panksepp (2012), and Ekman (2013).

3.2 The circumplex theory of core affect

Why discussing a theory of affect in the context of theories of emotion? First, because in practice, affect is often used interchangeably with emotion as a concept. Second, because of its historical relevance for the study of emotion. And third, because in many machine learning applications emotion is operationalized as core affect (Mollahosseini, Hasani, & Mahoor, 2019; Li et al., 2019; Kervadec, Vielzeuf, Pateux, Lechervy, & Jurie, 2018; Vielzeuf, Kervadec, Pateux, & Jurie, 2018)

The circumplex theory of affect was introduced by James Russell (1980). However, it has direct precedent on the Harold Schlosberg's *three dimension theory of emotion* (Schlosberg, 1954). In short, Schlosberg proposed that emotion can be represented by three independent continuous dimensions: *pleasantness-unpleasantness*, *attention-rejection*, and *tension-sleep* (i.e., activation). Russell's circumplex theory carried on with this tradition with two differences: first, Russell theorized about *affect* instead of emotion; second, he proposed affect can be represented by two dimensions instead of three. Russell has encapsulated his views on affect under the rubric of *core affect* (Russell & Barrett, 1999; Russell, 2003). I mentioned Russell's definition in our discussion about *what emotions are not*, but I reiterate here for completeness:

"A neurophysiological state that is consciously accessible as a simple, nonreflective feeling that is an integral blend of hedonic (pleasure–displeasure) and arousal (sleepy–activated) values" (Russell, 2003, p., 147). From the *nonreflective* nature of core affect we can derive a couple of interesting properties: first, core affect can be experienced without reference to any *object*. This is, it does not need to be directed towards anything (although it can be, but this is incidental to its definition). For instance, people experience "love" towards *someone* or *something*, it has an *object*, but core effect does not need to: people can experience free-floating anxiety or calmness without attaching such feeling to nothing; second, core affect can be experienced without having to be named, interpreted, or attributed to any cause, as no cognitive processing is demanded (Russell, 2003; Russell & Barrett, 1999).

A *circumplex* is a pictorial representation composed of a Cartesian two-dimensional plane and a circle centered at the origin. Circumplex structures have been used in the literature to describe the correlation structure of several psychological constructs (Gifford & O'Connor, 1987; R. E. Plutchik & Conte, 1997; Stanislawski, 2019), hence are not unique to affect. In Russell's original work (1980), a group of 34 undergraduate students was asked to compare pairs of emotion-related words (e.g., tired, glad, afraid, calm), allowing Russell to form a similarity matrix based on similarity judgments. Multidimensional Scaling was used to assess the matrix and obtain the *geometric representation* (i.e., the so-called circumplex) of the associations among the emotion-related terms. From this analysis, Russell concluded that a two-dimensional bipolar space could represent the spacial relationship among terms. Such results have been replicated multiple times with a variety of factor analysis techniques (Russell, Lewicka, & Niit, 1989; Posner, Russell, & Peterson, 2005; Yik, Russell, & Steiger, 2011). *Fig. 1* shows a sketch of the prototypical configuration of a circumplex, with the horizontal axis representing *hedonic valence*, and the vertical axis representing *arousal*:

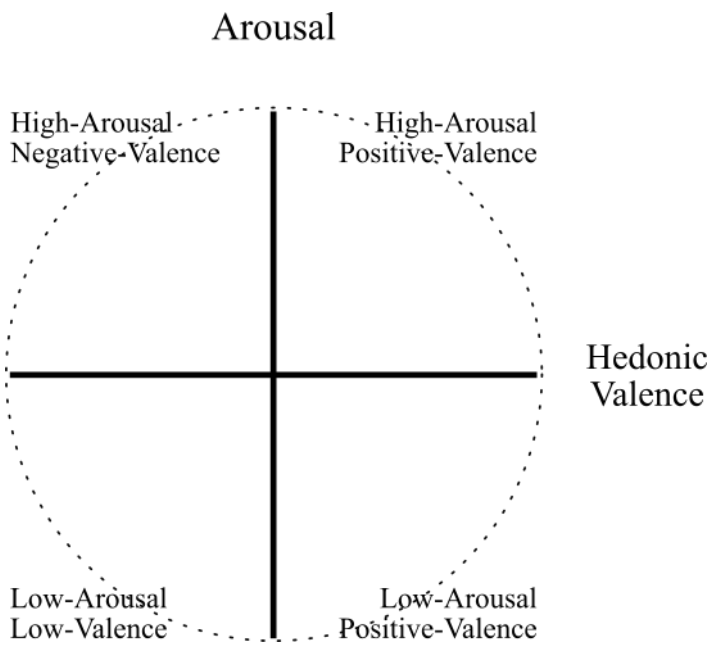

Figure 1. Circumplex model sketch.

As I mentioned earlier, core affect is considered to be a fundamental part of emotion, but it should not be confounded with it. About this distinction, Yik et al. (2011) have said: "Core Affect is a state accessible to consciousness as a single simple feeling (feeling good or bad, energized or enervated) that can vary from moment to moment and that is the heart of, but not the whole of, mood and emotion." (p., 705). Nonetheless, the fact that core affect is at the center of emotion has made it an increasingly common way to *operationalize* emotion in machine learning, particularly as an alternative to the discrete categories proposed by basic emotion theorists (Mollahosseini et al., 2019; Li et al., 2019; Kervadec et al., 2018; Vielzeuf et al., 2018).

The issue of the ontological status (i.e., natural kind status) of core affect is independent of the ontological status of emotion. Core affect theorists have argued that core affect can be considered an *irreducible primitive* component of emotional experience with roots in our evolutionary history (Barrett & Bliss-Moreau, 2009a; Russell, 2003) (although such presumption has been challenged (Scarantino, 2009)), more or less in the same way that basic emotion theorists have argued about the status of basic emotions like fear or anger (P. Ekman, 1992; Izard, 2007).

Nonetheless, regarding the status of emotion, core affect theorists have also argued that emotion *are not natural kinds*, regardless of the fact emotional experiences incorporate "emotion primitives" as hedonic valence and arousal state in their construction (Yik et al., 2011; Russell & Barrett, 1999).

3.3 The functionalist theory of emotion

Throughout human history, emotions have often been characterized as *forces of disruption* by many cultures: in the Buddhist tradition, *Tanhā*, a Pāli word meaning *thirst*, *craving* and/or *desire*, is regarded as the fundamental source of human suffering (Harvey, 2012); the Stoics sustained that emotions arose from incorrect judgement, and once an individual attains moral and intellectual perfection, it would not engage in such mistakes anymore (Sorabji, 2000); and the Christian Doctrine, regards emotions like *envy*, *wrath*, and *lust* among the seven deadly sins (Manning, 1874). In sharp contrast with this line of thinking, the *functionalist* theory on emotion see emotions as fundamentally *beneficial forces*, which facilitate adaptation to our natural and social environments, even "unpleasant" emotions as anger and fear (Adolphs, Mlodinow, & Barrett, 2019; Anderson & Adolphs, 2014; Averill, 1980; Keltner & Gross, 1999).

Broadly speaking, functionalist views on emotion can be divided into the ones that emphasize their *biological basis* and *evolutionary origin* (Adolphs et al., 2019; Anderson & Adolphs, 2014; Keltner & Gross, 1999) and the ones the that emphasize their *social goals* (Averill, 1980; Gordon, 1991). An example of the former is Keltner & Gross' (1999) view of emotion as "an episodic, relatively short-term, biologically based patterns of perception, experience, physiology, action, and communication that occur in response to specific physical and social challenges and opportunities." (p. 468). An example of the latter is Averill's (1980) definition of emotion as "a transitory social role (a socially constituted syndrome) that includes an individual's appraisal of the situation and that is interpreted as a passion rather than as an action" (p. 312). The commonality among these variants, is the change of focus from the *what* emotions are to the *why* emotions exist. In this sense, emotions are defined in terms of the *purpose* they serve,

this is, how they facilitate adjustment to the challenges and opportunities in the environment, rather than by how they are represented in the mind or the specifics of the organism response.

Consider Adolph's (2013) perspective about *fear*: functionally, *fear* main purpose is to mediate the relationship between *threat-related* stimuli, and adaptive responses to cope with that threat. Threatening stimuli can vary: a predator, a cliff, a snake, or even a math test, all can trigger *fear*. Adaptive responses can also vary: fighting, avoidance, or freezing, all can be functional responses to threat. Hence, although there may be stereotypical stimulus-response patterns related to fear within species, ultimately, is *the functional purposes of emotion what defines its nature*. Accordingly, whatever it was an individual felt that lead them to cope with a threatening stimulus, can be cataloged as fear.

From an evolutionary perspective, basic emotion theory can be cataloged as functional as well: the proposal is that basic emotions emerged precisely because of their functional role in facing challenges and opportunities in our physical and social environments (Izard, 2007; Panksepp & Biven, 2012). Nonetheless, basic emotion and functional perspectives diverge in that for the former, evolution generated a precise set of emotion clusters (e.g., happiness, anger, fear, etc) that can be recognized by their signature physiological responses and facial expressions (P. Ekman & Keltner, 1997) , whereas for the latter, the physiological responses and face configuration are less stereotyped, which makes them significantly harder to identify (Adolphs & Anderson, 2018).

The fact that stimulus and responses related to emotions can vary may lead some to think that there is no way to scientifically pinpoint specific emotions. Yet, according to this view, this is not necessarily the case, and significant effort has been made by researchers to identify the plethora of neurobiological and behavioral patterns associated with emotions (Adolphs & Anderson, 2018). For instance, it has been proposed that the amygdala is implicated in the recognition, expression, and experience of fear (Adolphs, 2013; J. LeDoux, 2003; Tovote, Fadok, & Lüthi, 2015). Such findings have motivated press coverage of the amygdala as the "brain's fear center" (Bhanoo, 2010), which in turn have motivated some researcher to publish pieces rejecting that interpretation (J. E. LeDoux, 2014; Barrett, 2017a). The point is not

that the amygdala is not implicated in fear, but that it's part of a larger and complex set of neurobiological and behavioral responses associated with it (Adolphs & Anderson, 2018; Adolphs, 2019).

3.4 The theory of constructed emotion

In 1543, Nicolaus Copernicus published "*De revolutionibus orbium coelestium*", introducing the radical idea that the Earth is the one orbiting around the Sun, instead of the Sun around the Earth. Such a notion contradicts everyday experience that suggests the opposite is true: when we look up to the sky, it does not look or feel like *we* are moving around the Sun, but like the Sun "rises" and "sets" around *us*. Similarly, the constructionist view on emotion suggests that we have *fundamentally misunderstood the nature of emotion*. Although everyday experience may suggest emotions are things that happen *to us*, as some form of internal "force" waiting to be activated in our brains and bodies, on the contrary, emotions would be ad hoc mental categories created *by us* in response to some meaningful event (Barrett, 2017a, 2017b). This implies that no two emotional episodes are identical, as the factors bringing them into existence, *your past experiences* (crystallized in your brain and body) plus *your present circumstances*, are unique. From this uniqueness it follows that variation happens not only *between* individuals and emotional episodes, but also *within* individuals and emotional episodes (Barrett, 2013): individuals differ not only in how they experiment happiness, but instances of happiness can vary within the same individual. Going further, the whole concept of emotion can be completely different for different individuals. The "constructionist" character of this perspective derives precisely from this: emotions as complex *human fabrications*, instead of naturally occurring entities like fermions and bosons.

The constructionist perspective has been championed by Lisa Feldman Barrett and colleagues (Barrett, 2017a, 2017b; Barrett & Simmons, 2015; Barrett, Wilson-Mendenhall, & Barsalou, 2015), in sharp contrast with what Barrett identifies as the *classical view on emotion*. In short, the classical view of emotion presumes that emotions are distinct natural entities

(e.g., *fear*, *sadness*, *happiness*), that emerged over our evolutionary history for adaptive and functional purposes, and therefore, emotions must posses universal and objective biological markers to be discovered by emotion scientists (Barrett, 2017a). Basic emotion theories (P. Ekman & Cordaro, 2011; Tracy & Randles, 2011) and functionalist perspectives on emotion fits this description (Adolphs & Anderson, 2018; Anderson & Adolphs, 2014; Keltner & Gross, 1999).

The contrast between the constructionist and classical views derives from profound epistemological differences not only about the conceptualization of emotion but also about the nature of *psychological entities*. According to Barrett (2017a, 2017b), classical perspectives on emotion have adopted *folk psychology categories* from western culture and philosophy and culture, that do not offer adequate conceptual grounds for a contemporary scientific theory of emotion. In Barrett's view (2012), emotions are *socially constructed realities*. In other words, emotions are as real as *money* or *marriage*: as *human conventions* that acquire substance in so far we decide to attach meaning to them. Marriage does not grow on trees, but that does not make it less real. About this proposal, Barrett's has stated:

> "Humans create ontologically subjective categories to serve functions that help constitute social life. According to Searle, such functions are the glue that holds a human society together. If emotion categories are ontologically subjective categories, then they can be thought of as collective cognitive tools that allow members of the same culture (and even different cultures, depending on the categories, of course) to represent and shape the social meaning of physical events." (p. 419)

At the individual level, such ontologically subjective categories are created by a blend of mental representations and body reactions, meaning that emotions can be understood as "those embodied representations that shape the animal's action and become the animal's experience of the world in that upcoming moment" (Adolphs et al., 2019). In this sense, emotions are *active inference processes* enacted by the brain as needed on the basis of past experiences and the current circumstances (Barrett, 2017b). In short, emotions would be a mix o *category*

generation and *instance categorization* problem.

The "clashing" between classical and constructionist views about emotion is not new. It can be thought as particular instance of long-standing debates in philosophy and science: *platonism* versus *aristotelianism* (Ariew, 2002), *nominalism* versus *realism* (Armstrong, 1978), and *essentialism* versus *social constructionism* (Raskin, 2002; Sayer, 1997). It is not unique within psychology either. For instance, essentialism versus social constructionism has been an important point of debate in the study of human sexuality (DeLamater & Hyde, 1998) and gender (Bohan, 1993).

The constructionist view on emotion contests classical views not only on conceptual grounds but in empirical ones too. Barrett and colleagues have discussed extensively about the evidence supporting their views (Barrett et al., 2007; Barrett, 2017a; Barrett & Satpute, 2019; Barrett et al., 2019). I will not review such an extensive corpus of empirical evidence, as our focus is conceptual exposition and analysis rather than to recapitulate the empirical evidence. In a nutshell, the issue revolves around the evidence (or lack of evidence) regarding the existence of "blueprints" and/or "biological markers" of emotion, ranging from brain circuits for fear or anxiety as proposed by functionalist views (Adolphs, 2013), to facial expressions as proposed by basic emotion views (P. Ekman, 1992; Tracy & Randles, 2011). Barrett and colleagues have suggested that the evidence supporting the interpretations of classical views is lacking or inconclusive at best (Barrett, 2017a; Barrett et al., 2019). The constructed theory of emotion emerged precisely as a way to capture the, according to Barrett and colleagues, the main lesson of decades of research on human emotions: that *variation is the norm rather than the exception*, in all aspects related to emotion, between cultures, between individuals, within individuals, and even within emotional instances (Barrett, 2017a).

4 The measure of emotion: establishing emotion ground truth

Considering the variety of perspectives on emotion, and that such views disagree about several fundamental aspects, it is fair to ask what has psychology to offer to machine learning researchers to guide the development of affective computing systems and technology. At first glance, the conceptual landscape of emotion looks messy and contradictory, which makes it hard to decide how to operationalize emotion in applied settings. In our view, a productive approach is to consider what does entail to adopt different perspectives on emotion, in other words, what kind of theoretical compromises and practical effects are tied to different views on emotion. For this analysis, I will follow Barrett's distinction between *classical (or common) views* and *constructionist views* on emotion (Barrett, 2017a). Admittedly, this simplification significantly facilitates the analysis, but comes at the cost of introducing some degree of misrepresentation, particularly for classical views, which contain a considerable degree of variation. I ask the reader to keep this in mind while pondering the relative merits of each perspective.

A first important clarification is the duality between recognizing emotions in *others* and recognizing emotions in *oneself*. Until this point, I have only considered what emotions are in general. However, in practice, many machine learning applications focus on emotion recognition, particularly with supervised learning techniques which aim to correctly classify emotions based on mainly four modalities: (1) *images* (e.g., Mollahosseini et al., 2019), (2) *text* (e.g., Seyeditabari, Tabari, & Zadrozny, 2018), (3) *video* (e.g., Bargal, Barsoum, Ferrer, & Zhang, 2016), (4) and *audio signals* (Khanna, Davletcharova). To analyze the implications of different perspectives about emotion for every modality would require a separate article, therefore I will focus our attention on the case of recognizing emotions from images and video. I selected these two since both involve visual input which makes their analysis closely related. I also believe my analysis can be, for the most part, extended to other modalities without significant alteration.

As supervised learning algorithms require labeled datasets, the first conundrum for machine learning researchers is how to obtain such labels: a first option entails asking *judges* to categorize

(with discrete or continuous approaches) emotions in others, for instance, by looking at pictures of other people faces (e.g., Mollahosseini et al., 2019); a second option is asking people to *self-report* their emotions in some sort of controlled setting, for instance, by annotating their emotions when writing a brief emotional note (e.g., Kleinberg, van der Vegt, & Mozes, 2020). Regardless of the method, once the labels are recorded, they are taken as *ground truth* for training a model. This is a critical decision, as it can validate or invalidate any claim regarding the capacity of the system to recognize emotions.

4.1 Recognizing emotions in others

Consider the alternative of asking judges to categorize emotions in others by looking at pictures of their faces: its validity relies *entirely* on the assumption that emotions in others can be accurately identified by looking at pictures of their facial expressions. From a classical view, particularly basic emotion theories, this is a completely valid procedure. Even more, *it has to be possible*, otherwise, the validity of this approach would be at risk. Ekman and colleagues (P. Ekman & Friesen, 1978; P. Ekman, Friesen, & Hager, 2002) have even developed the so-called *Facial Action Coding System* (FACS), a method for describing facial movements that breaks down facial expressions into individual facial muscle movements called *Action Units* (AU), which can be used for automated emotion recognition (e.g., Lien, Kanade, Cohn, & Li, 1998; Kapoor, Qi, & Picard, 2003). The hypothesized universality of facial expression should make it possible to recognize emotions with high accuracy *regardless* of context, culture, or individual differences. Therefore, having judges to label datasets of facial expressions is perfectly valid. On the other hand, when considering this procedure from a *constructionist perspective*, the exercise of making judgments about emotions based solely on pictures of faces as *ground truth* makes little sense. According to his view, there is no such thing as "universal facial expression of emotion", as facial configurations will vary *between* and *within* individuals (Barrett et al., 2019). Context, culture, and individual differences are fundamental pieces of the puzzle of inferring emotions in others. Further, even taking all that into account, different observers can make

substantially different judgments regarding emotions in others, since their own past experiences with emotion events will condition their inferences about emotions. Thus, the conundrum here boils down to the validity of the "universality of facial expression assumption". Although giving an answer to this question is beyond our scope, there is substantial evidence contradicting such assumption (Barrett, 2017a; Barrett et al., 2019; Gendron, Crivelli, & Barrett, 2018; Gendron et al., 2020).

As with images, recognizing emotions in others based on video sequences is no different from a classical emotion view: as long as certain facial configuration and physiological responses are present in the sequence, human judges or automated systems should be able to recognize emotions with reasonable accuracy, and importantly, such labels can be taken as *ground truth*. From a constructionist viewpoint, depending on the type of video sequence, recognizing emotions in others becomes more plausible. Recordings just adding movement to an otherwise contextless and artificial scenario are unlikely to improve the validity of the procedure. On the other side of the spectrum, highly naturalistic video recordings that incorporate context, audio, body language, plus other cues, should significantly improve the *plausibility* of correctly inferring emotions in others, just as it may happen in real life.

In either case, images or video, there is a fundamental difference that persists: from a *classical viewpoint*, the labels generated by judges or automated systems can, *in principle*, precisely match the emotion experienced by the individual in the image or video sequence based on objective biological markers; from a *constructionist perspective*, the chances of correctly labeling emotions experienced by others should increase as the information accessible by the judge is richer and more naturalistic, yet, *there is no way to be certain about the judgment correctness* other than asking to the individual in the image or video, a situation in which having judges becomes pointless. Further, since constructionist views do not limit the emotions that can be experienced to a predefined set of universal emotions, individuals can (in principle) generate an *unbounded* number of emotion categories. Therefore, a judge may devise an emotion category that is not even part of the emotion vocabulary of the person in the image or video. Moreover, the larger the cultural distance between the "judge" and the "judged", the higher

the chance of a mismatch in the emotion lexicon. This line of reasoning is supported by studies showing that when participants are allowed to freely categorize their emotions, they generate more emotion categories than what is proposed by classical views (Barrett et al., 2019; Cowen & Keltner, 2017; Hoemann, Crittenden, et al., 2019; Gendron et al., 2018; Wang, Lü, Zhang, & Surina, 2014). For instance, Cowen and Keltner (2017) asked a group of observers to categorize their emotional responses to 2,185 emotionally evocative short videos, with 9 observers for each video. Collectively, the observers generated 600 emotion categories with labels as diverse as "patriotism", "gratitude", "paranoia", and "divine inspiration".

The potential mismatch between *experienced* emotion and *perception* of emotion in others, raise additional issues regarding fairness and discrimination (Binns, 2017; Corbett-Davies & Goel, 2018) in emotion recognition systems. If, as constructionist theorist propose, emotions can not be reliably identified by judges looking at pictures of faces (or some other modalities), judges personal identity and biases becomes critical. In practice, machine learning systems would be learning from whatever biases or prejudice judges possess towards other groups, *automating* and *amplifying* those biases (O'neil, 2016). Minorities and marginalized groups in society are at greater risk to be perceived as displaying emotions often regarded as "negative", "inappropriate ", or even "threatening". For instance, it has been documented that black people, particularly black men, are perceived as more "threatening" than white people (Hester & Gray, 2018; Trawalter, Todd, Baird, & Richeson, 2008), and that women are perceived as more "emotional" than men in western cultures (Barrett & Bliss-Moreau, 2009b; Brescoll, 2016; Shields, 2013; Shields & Shields, 2002). If emotion recognition systems with this kind of biases are deployed in high-stake situations like hiring decisions, court decisions, or targeted policing, they can reinforce racial, gender, class, sexual orientation, and other types of stereotypes with harmful consequences for historically marginalized groups.

4.2 Recognizing emotions in oneself

At first glance, asking people to self-report their emotions in response to any stimulus may seem like an ideal situation to collect the so-called *ground truth* about emotion. Nonetheless, from a classical perspective, it is not obvious that self-report is better than judge assessment or machine-based assessment. People may lack the self-awareness or appropriated emotional lexicon to "correctly" report their emotions. Essentially, classical views regard emotion similarly to how physicians treat medical conditions: for instance, people may believe they suffer from COVID-19, but ultimately, physicians are the ones with the knowledge and tools (e.g., PCR test) to make the better judgment. It can be argued that people can be helped to do the "correct" self-assessment by constraining their responses to the "valid" emotion categories, this is, requiring people to categorize their emotions as one of the basic emotion classes of *fear*, *anger*, and so on. Indeed, this is a common practice in both psychology (Barrett et al., 2019; Gendron et al., 2018) and machine learning (e.g., Kleinberg et al., 2020). Ekman's original study utilized such approach (P. Ekman & Friesen, 1971), and more recently, several datasets in machine learning have been constructed in a similar manner (e.g., Barros et al., 2018; Gong, Huang, Wang, & Luo, 2011; Kossaifi, Tzimiropoulos, Todorovic, & Pantic, 2017; McDuff et al., 2013). For instance, Kleinber, van der Vegt, and Mozes (2020) have claimed to introduce the first *ground truth* dataset of emotional responses to COVID-19. In their study, they asked participants to write a few sentences about how they feel regarding the COVID situation, and then to choose which of *eight predefined emotions* better represented their feelings. In our view, this type of force choice categorization is deeply flawed as it makes it impossible to report any emotion that has not been preselected by the researcher, meaning *it is not falsifiable* (Popper, 2005). Even if researchers endorse basic emotion views, they still have to deal with the choice of the subset of basic emotions to incorporate, as there is no consensus regarding which emotions count as basic (Tracy & Randles, 2011), i.e., a significant degree of *arbitrariness is unavoidable*. This is not to say that self-report is not interesting or useful in the study of emotion from this perspective, but it is not an "objective" way to establish the ground truth for an emotion

recognition system.

Constructionist perspectives, again, contrast with classical views by taking the opposite approach: self-report of emotion is indeed *better* than observers judgment for the purpose of establishing emotion ground truth. This is a logical consequence of the constructionist way of understanding emotion: as emotions are constructed by individuals on the spot based on internal and external cues, there is no better way to establish the *ground truth*. Judges with a rich background and contextual information may approximate the *ground truth*, but the ultimate judge is the person experiencing the emotion. In this sense, emotion categorization is more like food tasting than medical evaluation: the chef may claim their food is objectively tasty, but the taster is the ultimate judge of the food tastiness. "Taste" is simply *not an objective attribute of the food*, but a subjective experience constructed by the individual, in the same manner that *emotion* is not an attribute of an event, but an artifact of the human mind and body in anticipation or response to it.

The constructionist approach poses significant challenges to establishing the ground truth for the task of building emotion recognition systems: it is not only the case that self-report is better but researchers would also want to elicit emotion as realistically as possible, which has proven to be a remarkably difficult challenge in psychology (Gilbert & Malone, 1995; Levitt & List, 2007; Pearl & Bareinboim, 2014). Moreover, they have to strictly separate systems that *predict labels generated by judges* (i.e., other people assessments of emotion), from systems that *predict labels generated by individuals experiencing the emotional event*, i.e., self-report. Strictly speaking, from this perspective, authors claiming to have built an emotion recognition system based on labels generated by judges, actually have built a system that *predicts observers assessments of emotions in others*. In other words, many emotion recognition systems, just tell us how certain picture or video would be categorized *by the judges that provided the labels*, not the emotion as it was experienced by the individual in the picture or video. This situation, as I have discussed in the previous section, may have severe ethical consequences for the deployment of emotion recognition systems.

In sum, from a classical perspective, scientists and their tools are the ultimate judges of

emotional experiences, whereas from a constructionist perspective, individuals are the ultimate judges of their own emotions. Thus, ML/AI researchers should be wary about their choices when deciding on how to establish emotion ground truth for their systems.

5 Conceptual approaches about the study of emotion in machine learning

5.1 The discrete vs continuous emotions narrative

My reading of the literature about emotion in machine learning, reveals that researchers tend to portray the theoretical landscape of emotion in psychology as a debate between two camps: on the one hand, the *Discrete Emotion Theorists* who sustain that emotions are better described as distinct categories like Ekman's (P. Ekman & Cordaro, 2011) basic emotions, or as combinations of "primary emotions" like the ones proposed by Plutchik's wheel of emotions (R. Plutchik, 2001); on the other hand, the *Continuous Emotion Theorists* as James Russell and Albert Mehrabian (1977), who sustain that emotions are better described as two or three orthogonal dimensions and that discrete emotions can be constructed as linear combinations of such components. The former perspective would lend itself better for multigroup classification approaches, whereas the latter for regression-based approaches. Thus, the chasm dividing emotion theorists in psychology, according to this narrative, would be whether emotions are better described as *discrete categories* or as a couple of *orthogonal dimensions*. Examples of this narrative can be found in (Seyeditabari et al., 2018), (Thanapattheerakul, Mao, Amoranto, & Chan, 2018), (Shu et al., 2018), (Poria, Cambria, Bajpai, & Hussain, 2017), (Gunes et al., 2011), (Kervadec et al., 2018), (Vielzeuf et al., 2018), (Sethu et al., 2019), and (Devillers, Vidrascu, & Lamel, 2005). Even though the discrete versus continuous dichotomy is a common narrative, some authors have offered a more complex account of emotion theory in psychology (e.g., Cowie et al., 2001; Fragopanagos & Taylor, 2005).

Although is true that the discrete versus continuous representation has been a subject

of debate, in our view, the aforementioned characterization does not fully capture the main points of controversy surrounding emotion theory in psychology (see Section 3 on "Theories of emotion"). Indeed, classical views on emotion sustain that emotions are discrete entities, but it is not the case that Russell and others have argued that emotions *must be* characterized as combinations of the continuous dimensions of valence and arousal, and nothing else. As I have discussed in previous sections, Russell's circumplex model theorizes about *affect*, not emotion, and Russell himself has stated that emotions can not be reduced to affective states as *valence* and *arousal* (Russell & Barrett, 1999; Russell, 2003). Similarly, constructionist theorists as Barrett have been less concerned about whether emotions can be represented as discrete categories or orthogonal dimensions, and more with whether emotions are natural kinds (Barrett, 2006), universal across individuals and cultures (Barrett et al., 2019), and about the mechanism that allows humans to generate emotion concepts (Barrett & Simmons, 2015; Barrett, 2017a). Importantly, constructionist views on emotion do disagree with the notion that emotions should be described as a predefined set of discrete categories (i.e., basic emotions or similar) but it does not negate the idea of humans generating situated *ad hoc categories* to describe their emotional experiences (Hoemann, Xu, & Barrett, 2019; Hoemann, Devlin, & Barrett, 2020). The crux of the matter is that the kind of subjective categories individuals generate, according to constructionist views, are not like the ones described by classical emotion theorists. On the contrary, the kind of categories people generate are situated, flexible, learned, variable, and a blend of things that can be described in a dimensional fashion like valence, and things that may be understood more or less as categorical, although in a qualitatively different fashion than classical views propose.

Regardless of the chosen perspective, I sustain the "discrete versus continuous emotion" narrative is insufficient as a characterization of the theoretical landscape in the psychology of emotion, severely limiting the pool of operationalization alternatives for machine learning researchers.

5.2 Examples of conceptual approaches in machine learning research of emotion

The machine learning research landscape on emotion is vast hence trying to summarize the empirical evidence for each approach would be futile and well beyond the scope of a conceptual review. Nonetheless, examining a couple of examples from each of the main perspectives will help to give substance to our discussion and clarify some claims.

5.2.1 Discrete emotions examples

The *Static Facial Expressions in the Wild* (SFEW) database, introduced by Dhall, Goecke, Lucey, and Gedeon almost a decade ago (2011), is one of many databases that implement a discrete approach to emotion classification. Since then, it has been utilized in numerous publications to train emotion recognition system in the context of the *Emotion Recognition in the Wild Grand Challenge* (Dhall, Goecke, Joshi, Wagner, & Gedeon, 2013), which is held every year at the *International Conference on Multimodal Interaction*. The SFEW corpus contains 700 face images extracted from 37 movies, labelled for six basic emotions (i.e., *anger*, *disgust*, *fear*, *sadness*, *happiness*, and *surprise*) plus a *neutral* expression category. The labeling process was done for two independent trained raters. This approach is sometimes called *pseudo ground truth*, as it relies on trained raters rather than directly questioning the individual experiencing the emotional event. In practice, the labeling strategy implies that any model trained on this dataset is predicting the *perceptions about others' emotions by the two anonymous labellers*, not the emotion of the individual in the image. There is nothing wrong with training models for such a purpose in itself, the issue arises when is claimed to have trained a model to recognize the emotions of the individuals in the images *in general*, which is unwarranted. The second issue is the assumption that the only perceivable emotions are the six basic emotions of *anger*, *disgust*, *fear*, *sadness*, *happiness*, and *surprise*, and the neutral expressions, meaning the labellers have no alternative but to "perceive" those six emotions, making emotion attribution from he model even more questionable.

In a recent article, Kleinberg, van der Vegt, and Mozes (2020) introduced the *Real Word Worry Dataset* (RWWD) consisting of 5,000 texts about emotional responses to COVID-19. Since each text was annotated by the individual that wrote it, the authors describe it as the first *ground truth* dataset of emotional responses about the pandemic. However, a closer look at the methodology reveals that this is a questionable claim. The authors asked 2,500 participants to write a long (min. 500 characters) and a short text (max. 240 characters) regarding how they feel about the coronavirus situation. The participants also had to select which of eight predefined basic emotions (i.e., *anger*, *anxiety*, *desire*, *disgust*, *fear*, *happiness*, *relaxation*, and *sadness*) best captured their feelings in that moment, than then were used as ground truth for correlational and predictive modeling. Although the RWWD overcomes the trained annotator annotator problem, the second problem remains: the set of emotions participants could express were preselected for them *by the researchers*, with defeats the purpose of establishing the ground truth of participants' emotions. Following this approach can only lead to a self-fulling prophecy as there is no way to prove the authors wrong about their claim of measuring people's emotions.

5.2.2 Continuous emotions examples

The *Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression dataset* (SEMAINE), is a popular source for modeling emotion in machine learning (McKeown, Valstar, Cowie, Pantic, & Schroder, 2012). It contains multi-modal audio-visual recordings of interactions between an "operator" (either an artificial agent or a human) and a "user" (always a human). Two modalities were utilized for constructing SEMAINE: the *SAL* modality where the "operator" role was played by an artificial agent, and the *PowerPoint SAL* modality where a human emulated the artificial agent behavior. The interactions were modeled after TV chat shows to trigger emotionally charged responses from human users.

There are several interesting facts about the annotation process for SEMAINE. First, annotations were made for five continuous emotion dimensions and 27 discrete emotion categories. The continuous emotion dimensions were: *valence*, *activation*, *power*, *anticipation/expectation*, and *intensity*, which were chosen based on a influential study by J. Fontaine et al (2007) (*power*

was an additional dimension incorporated by the authors). Second, the annotation was done by 6-8 raters per clip, but of unknown demographics. Third, study participants were mostly caucasian undergraduate and graduate students from eight different countries with 62% female and 38% male participation. This approach certainly provides richer information about emotion than most studies, yet the data still contains important limitations *if* the goal is to build emotion recognition and synthesis systems: trained raters (i.e., "pseudo-ground-truth"), narrow participant's demographics, and unknown rater's demographics. It is also important to pay attention to how the J. Fontaine et al. study validated their four-dimensional model: by asking a group of 531 British, Swiss, and Dutch young adults to rate 24 hand-picked emotion-related words in 144 features selected by the researchers. This is not to say such dimensions are necessarily "invalid", but it should make researchers wary of the limitations imposed by the sociocultural specificity of such approach.

Another example is the *AffectNet Database*, introduced by Mollahosseini, Hasani, and Mahoor (2019), which is probably the largest corpus of images for emotion recognition and modeling available. Its defining feature is their 1,000,000 face images collected "in the wild", this is, by querying 1250 emotion-related keywords in six different languages directly from the internet. Each image was annotated for eight categorical labels and for the dimensions of *valence* and *arousal*. The continuous dimensions were annotated by 12 trained raters from unknown demographics. Although the issue of external annotators prevails, the sheer size of the dataset plus avoiding the use of posed photos with stereotypical face configurations is an advantage. Yet, some additional issues with this strategy must be noted: images from the internet are most likely *biased* towards certain subgroups of the world population with greater representation on the internet, and that most images probably come from actors in movies and TV shows, stock photos, or from a small subset of emotional expressions as people rarely upload to the internet pictures of themselves depicting the whole spectrum of their emotional lives.

6 Main issues in the conceptualization and measurement of emotion in machine learning

Ideally, the conceptualization and measurement of any psychological construct must align. Several of the issues noted in the previous section are not necessarily a problem depending on the theoretical assumptions made by the researcher. In general, basic emotion approaches entail significantly fewer restrictions than constructionist approaches. For instance, from a basic emotion perspective, it is not problematic to utilize judges' ratings of other people's emotions as the so-called *ground truth*. Nonetheless, is our opinion that utilizing such a lax standard does not reflect our current understanding of human emotions. Hence, I urge researchers in the ML/AI community to adhere to higher standards and consider the following issues seriously.

6.1 Perceived versus experienced emotion

Most databases I could find available employ a judges system to construct the labels for machine learning models of emotion. Meaning that, in practice, most databases of emotion are about *perceived emotion*, rather than *experienced emotion*. Such a situation becomes a problem only when we claim emotion recognition systems can identify how people are feeling (e.g., how the individual in a picture is feeling), rather than how a certain subset of the population (i.e., the judges) would perceive their emotions.

6.2 Sociocultural profile of judges and the subject of emotion

My reading of the literature indicates that the vast majority of available databases and studies rely on a narrow subset of the human population: caucasian young adults (i.e., college students) from western industrialized countries. In psychology, that type of group is commonly known as WEIRD population (Western, Educated, Industrialized, Rich, and Democratic) (Rad, Martingano, & Ginges, 2018). This is true for both the people whose emotions are being categorized and the judges performing the categorization. Even databases like AffecNet (Mollahosseini et

al., 2019) that sample images from "the wild" (the internet), use raters of such characteristics to construct the labels.

It can be argued that utilizing WEIRD population to build databases of human emotions is not a problem as long the models are deployed in the cultural context they were built. I believe this is shortsighted for several reasons. First, even within WEIRD populations there exist significant variation (e.g.: gender, political ideology, language, religion, etc.) which is not being accounted for. Second, it is highly unlikely that once the databases are made available, researchers, companies, and governments will use them *only* in the sociocultural context of creation. Third, the widespread adoption of systems based on highly biased data can do *significant harm*, particularly to vulnerable populations. Misreading the emotions of a person of color during a job interview can further reproduce social inequalities in job opportunities. It has been reported in the press that companies like *Affectiva* are already offering that type of services to large employers like *Unilever* and *Dunkin Donuts* (Zetlin, 2018), which should be a matter of concern.

6.3 Naturalistic versus non-naturalistic scenarios

Eliciting emotional responses from human participants is a challenge. A common practice is having actors posing facial expressions in accordance with western stereotypes for basic emotions, which is an example of a *non-naturalistic* database. Alternatively, researchers can collect large samples of photos, videos, audio, or text from the internet, which is a more *naturalistic* approach. Nevertheless, the "naturalistic" character of information collected from the internet is up to debate, as the information available online is a mix of fabricated content for mass consumption (e.g.: movies, news articles, video blogs, etc.), and content that people put online as part of their daily activities (e.g.: tweets, *Facebook* post, *Instagram* photos, etc.). Even when the data comes exclusively from sources like *Instagram*, which presumably contain a higher proportion of information coming from people's daily activities, is still likely such information is biased as people tend to carefully choose what images of themselves to make public. For

many corporations working on emotion recognition and synthesis software, this might or might not be a problem, as their goals are exclusively to deploy systems that can perform well on the Internet. Hence, they are sampling from the same source the same place they are deploying. Yet, researchers must be aware that the Internet is no guarantee for "naturally" occurring emotional expressions.

6.4 Predefined versus open-ended emotion categories/dimensions

A major issue for emotion databases is the practice of pre-selecting the "valid" emotion categories or dimensions by researchers. This is true for both studies utilizing self-report and studies utilizing trained judges. Put simply, this means that researchers decide what people can feel or judges can perceive in their studies. Leaving aside the micro-politics aspect of this practice, it is highly problematic as significantly distorts the measurement of people's emotions in any setting. It can be argued that if the study focuses exclusively on basic emotions, this is valid. In our view, too much evidence has been collected contradicting the idea that such a narrow set of categories can encapsulate the whole of human emotions (Barrett, 2017a; Barrett et al., 2019) to make the approach justifiable. Further, there is substantial evidence that when people are asked to freely label their emotions they come up with dozens or even hundreds of categories (Barrett et al., 2019; Cowen & Keltner, 2017; Hoemann, Crittenden, et al., 2019; Gendron et al., 2018; Wang et al., 2014).

6.5 Self-report is cultural and individual specific

People sometimes use the same words to signify different things, even within the same culture, even within the same family. We can go even further, and say that the *same person* can mean different things when utilizing the same word to describe its feelings on different circumstances. The author is this article is sure that although he says he is "happy" after both a good meal and after LeBron James won his 4th NBA Championship, those two instances of "happiness" mean different things, and above all, feel very differently. Although some researchers may think self-

report is the solution to obtaining the so-called ground truth about emotion (e.g.: (Kleinberg et al., 2020)), it is not, and it is better to be aware of such limitation: *self-report of emotion us cultural and individual specific.*

6.6 The purpose and uses of databases created by others matter

When examining the studies that originally collected several of the major databases for human emotions, it becomes clear that in most cases researchers had very specific uses in mind. For instance, SEMAINE (McKeown et al., 2012) was created with the purpose of analyzing emotion in conversational situations and to build artificial systems able to recognize social signals from humans, whereas RWWD was created to model emotional responses to the COVID-19 pandemic and for topic modeling (Kleinberg et al., 2020). Yet, once the databases popularize and become freely available, it is easy for researchers to use such data for goals that greatly differ from their original purposes. I do not believe this is a problem in itself as in many cases researchers can and do take the appropriate safeguards to make responsible use of the data. The same can be said for already trained models, which then are used for prediction of data that was not part of the original study. The website `https://paperswithcode.com` contains hundreds of machine learning articles with code precisely for this purpose (for emotion recognition examples see `https://paperswithcode.com/task/emotion-classification`). That being said, there are clear risks in taking data from secondary sources for new studies. The clearest in our opinion is the one I mentioned already: to use databases constructed for and by WEIRD populations that are taken out of context to synthesize or classify emotions of people from different sociocultural groups. of course, this is not the only risk but is one of the major ones given its potential economic, social, and political repercussions.

7 Avenues for improvement and innovation in the study of emotion in machine learning

Perfectly measuring emotion to construct an ideal dataset it is not only nearly impossible but it might also be counterproductive. As with most phenomena in science, a better approach is to do incremental progress with an ideal situation as a *guide* rather than as a hard constrain. In this spirit, I aim to identify several avenues for improvement in the conceptualization and measurement of emotion in machine learning. Many of these ideas derive directly from previous sections, yet I believe it will be productive to organize such ideas for clarity of thought.

7.1 Recognize human emotion's complexity and fuzziness

Is our opinion that researchers would benefit from recognizing the overwhelming complexity of human emotions. By recognizing I mean to abandon simplistic *operationalizations* of emotion, like asking two or three raters to label complete datasets with only five or six so-called basic emotions categories. Even if a researcher strongly endorses discrete measurement approaches, there are plenty of richer and more complex alternatives like the one introduced by Cowen and Keltner (2017). The recognition of such complexity should not only change the practices surrounding the measurement of emotion but also how results are communicated to the academic and non-academic audiences: as small yet significant steps in the right direction rather than as a "solved problem" with equal or better performance than humans judges, which is a common message in the media nowadays.

7.2 Self-report is better than the alternatives

Self-report has several shortcomings as I have mentioned in the previous section but still, I propose is a better alternative to the perception of a couple of judges (no matter how well trained they are), and certainly better than automated systems like FACS (e.g., Lien et al., 1998; Kapoor et al., 2003). At a minimum, self-report directly question the actual source of

the emotional event rather than relying on the opinion of a trained annotator with little to non-first-hand knowledge about the culture and identity of the judged individual. This is of course contingent upon the goal of the study, as we are implicitly assuming researchers want to build systems for classifying or synthesizing *someone's emotions*, rather than systems for classifying *someone else's perceptions* of emotions.

7.3 To utilize both discrete and continuous approaches for measuring emotion

The *discrete* versus *continuous* approach to measuring emotions will probably continue to divide human emotions' researchers for years to come. In this context, it advisable to take an eclectic approach measuring emotions in a variety of manners. This is not an endorsement of any particular classification system of people's emotions. Instead, is a proposal to recognize the diversity and complexity of the elements that constitute and ultimately define human emotions. In other words, is key to be up-to-date with the latest developments on the conceptualization and measurement of emotions and to incorporate best practices from a variety of perspectives.

7.4 Emotions are more than a label

Although there are several systems for measuring emotions available, I believe researchers must capture as much *contingent information* as possible to better characterize human emotional expressions. A non-exhausting list would include facial expressions, body language, physiological responses, sociodemographic characteristics, affective state, cultural background, and context. What I am trying to convey is the idea that simply asking people to put a label to their feelings in a particular instance is not enough, as it is missing several elements that are also part of the construction of the emotional event. The more "features" of such a circumstance are captured, the richer the characterization of emotional events.

7.5 There is no such a thing as "cultureless" and "contextless" emotions

We must insist on this point: if culture and context are not taking into account, there is no way to tell the emotional meaning of words, body language, or facial expressions. People do not experience or perceive emotions in a cultural and situational vacuum. There is plenty of evidence supporting this notion (Gendron et al., 2020; Gendron, Roberson, van der Vyver, & Barrett, 2014; Gendron et al., 2018; Hoemann, Xu, & Barrett, 2019), yet I believe is easy to see for the casual observer that seemingly identical facial configuration or words can acquire strikingly different meanings in different situations and cultures. Eyebrows coming down and together, wide-opened eyes, and tightly pressed lips, can communicate *rage* after an insult from a fellow driver in a hurry, and *elation* after hitting a last-second three-point game-winner shot in the basketball court. We can go further, and say that there words and facial expressions that do not have any clear meaning at all outside their cultural and situational context. True, we can try to forcefully "fit" others cultures unique emotional concepts to ours but such an exercise is an attempt to distort reality for our convenience which defeats the purpose of any scientific study. For instance, *Weñagkün* is a Mapudungun word of exquisitely complex emotional meaning: a mix of the long-lasting feeling of sadness after the loss of a loved one, accompanied by a hard feeling of emptiness where the person continues living a normal life, but occasionally experiencing a heavy feeling of pain accompanied from physical illness and body aches, particularly in the stomach and intestine (Sales & Lucena, 2000). We may be tempted to translate this as *grief*, *mourning*, or *sadness*, but all of them would be incorrect: it is *Weñagkün*. Although difficult, there already exist some datasets that incorporate some degree of context like the *EMOTions In Contex* dataset (EMOTIC) (Kosti, Alvarez, Recasens, & Lapedriza, 2019), and the community would benefit from more and better datasets in this direction.

7.6 Real-time and naturalistic approaches

The more naturalistic the circumstances under which emotions are captured, the more likely such expressions of emotion will reflect what researchers expect. Emotions are a prime example of psychological phenomena closely attached to events, situations, and people, that have acquired affective meaning throughout a person's lifetime. Although many procedures have been envisioned to trigger emotions in artificial settings, today we have at our disposition wearable technology (e.g.: smartphones and smartwatches) which provides a great opportunity to capture emotion in real-time and under every-day circumstances. An example of this approach is Hoemann and colleagues' study (2020), where 52 participants completed 14 days of peripheral physiological monitoring (i.e., electrocardiogram, impedance cardiogram, and bodily movement and posture) along with periodic self-report of their affective state. Similar approaches can unlock better and richer measurements of emotional experiences for future studies.

7.7 Do not restrict people's emotions

Pre-selecting a shortlist of "valid" emotions for either self-report or judges annotation is a widespread practice. Is our opinion that such a practice must either stop or be accompanied by the alternative of freely categorizing emotions. This alternative may indeed lead to an "explosion" of dozens or hundreds of categories which can be hard to handle. To this, I answer the following: that this is the *true* nature of the problem of emotion categorization. For too long, researchers have assumed that emotions *must be* a small subset of latent construct shared across all individuals and that studies must do their best to capture them. My opinion is that emotions are not fundamentally different than other domains of categorization, where humans are constantly generating new concepts to attach meaning to recurrent circumstances. In other words, machine learning researcher would benefit from thinking on emotions as a *large-multiclass categorization problem* like classifying *animals*, *types of plants*, or *every-day objects*. In those domains, it seems to be perfectly acceptable for researchers to have to deal with hundreds if not thousands of categories, but for some reason, that seems unacceptable for emotions. There are many

examples of people's capacity to generate new concepts on popular culture that reaffirm my point. The *Dictionary of Obscure Sorrows* `https://www.dictionaryofobscuresorrows.com/` is a project by graphic designer and filmmaker John Koenig, which aims to capture many popular neologisms for emotions. Words like *midding* "v. intr. feeling the tranquil pleasure of being near a gathering but not quite in it", or *kuebiko* "n. a state of exhaustion inspired by an act of senseless violence, which forces you to revise your image of what can happen in this world" are just a pair of provoking examples. Thus, in my view, there is no away this fact without significantly misrepresenting people's emotional experiences.

8 Conclusions

Before asking whether machines *can* synthesize or recognize human emotions, researchers must ask themselves whether they are measuring what *they think* they are measuring. Based on my analysis, I maintain that we are still far from truly capturing the complexity of human emotions and significant work remains to be done to accomplish such a goal.

For too long, researchers have relied on relatively simple and easy to operationalize ideas about the nature of emotion, like the so-called Ekman's "basic emotions" (P. Ekman & Friesen, 1971; R. Ekman, 1997). I sustain researchers must take steps towards increasingly complex and richer conceptualizations of emotion, with the goal to better capture the nuances and diversity of human emotional experiences. Contemporary research on human emotions have systematically showed that emotions are characterized by their *variability*, *fluidity*, *generativity*, and *dependence upon* the individual, context, and culture.

I have identified several points of concern regarding the current conceptualization and measurement of emotion: (1) the tendency to conflate *perceived* versus *experienced* emotion; (2) the narrow and unrepresentative sociodemographic profile (i.e., WEIRD population) of both human raters and the people subject to the emotional event; (3) the primacy of non-naturalistic databases over naturalistic ones; (4) the widespread practice of utilizing predefined emotion categories or dimensions instead of allowing people to freely categorize their emotions; (5) the fact

that self-report is no guarantee of objectivity; (6) and the importance of taking into account the original purpose of databases for emotion recognition and synthesis.

Further, I have also identified a number of avenues for the improvement and innovation in the study of emotions: (1) to recognize and complexity and fuzziness of human emotions, which too often is ignored; (2) that although it contains many shortcomings, self-report is a better strategy that trained raters in most instances; (3) the importance of utilizing an eclectic approach for the measurement of emotion; (4) the fact emotions are more than a label, this is, that studies must aim to capture a plethora of concurrent events when measuring emotion like facial expressions, body language, physiological responses, sociodemographic characteristics, affective state, cultural background, and context; (5) that context and culture are absolutely critical and *unavoidable* for the understanding of emotion; (6) that real-time and naturalistic approaches are a better strategy for measuring emotions and that current technology can facilitate this type of measurement; (7) and that people always must be allowed to freely categorize their own emotional experiences, as it is the only way to capture their emotional lexicon and subjective experiences.

Making progress in this direction it is not just a matter of scientific rigor and technological advancement, but also of *ethics* and *justice*. Historically marginalized and vulnerable groups have been systematically *underrepresented* and *misrepresented* in affective computing research, a situation that may have (and it is probably already having) harmful consequences for those groups. Affective computing technology is a double-edged sword: it can significantly improve quality of life and lead to economic advancement, but also can *automate* and *amplify* inequalities, injustice, and discrimination in society.

References

Adolphs, R. (2013). The biology of fear. *Current biology*, *23*(2), R79–R93. (Publisher: Elsevier)

Adolphs, R. (2019). Emotions are functional states that cause feelings and behavior. In *The Nature of Emotion, Second Edition* (pp. 6–10). Oxford University Press, New York.

Adolphs, R., & Anderson, D. J. (2018). *The neuroscience of emotion: A new synthesis*. Princeton University Press.

Adolphs, R., Mlodinow, L., & Barrett, L. F. (2019, October). What is an emotion? *Current Biology*, *29*(20), R1060-R1064. doi: 10.1016/j.cub.2019.09.008

Anderson, D. J., & Adolphs, R. (2014). A framework for studying emotions across species. *Cell*, *157*(1), 187–200. (Publisher: Elsevier)

Ariew, A. (2002). Platonic and Aristotelian roots of teleological arguments. *Functions: New essays in the philosophy of psychology and biology*, 7–32. (Publisher: Oxford University Press, USA)

Armstrong, D. M. (1978). *Nominalism and Realism: Volume 1: Universals and Scientific Realism* (Vol. 1). CUP Archive.

Averill, J. R. (1980). A constructivist view of emotion. In *Theories of emotion* (pp. 305–339). Elsevier.

Bargal, S. A., Barsoum, E., Ferrer, C. C., & Zhang, C. (2016). Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (pp. 433–436).

Barrett, L. F. (2006, March). Are Emotions Natural Kinds? *Perspectives on Psychological Science*, *1*(1), 28–58. doi: 10.1111/j.1745-6916.2006.00003.x

Barrett, L. F. (2012). Emotions are real. *Emotion*, *12*(3), 413. (Publisher: American Psychological Association)

Barrett, L. F. (2013). Psychological construction: The Darwinian approach to the science of emotion. *Emotion review*, *5*(4), 379–389. (Publisher: Sage Publications Sage UK: London, England)

Barrett, L. F. (2017a). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.

Barrett, L. F. (2017b). The theory of constructed emotion: An active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, *12*(1), 1–23.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019, July). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. doi: 10.1177/1529100619832930

Barrett, L. F., & Bliss-Moreau, E. (2009a). Affect as a psychological primitive. *Advances in experimental social psychology*, *41*, 167–218. (Publisher: Elsevier)

Barrett, L. F., & Bliss-Moreau, E. (2009b). She's emotional. He's having a bad day: Attributional explanations for emotion stereotypes. *Emotion*, *9*(5), 649. (Publisher: American Psychological Association)

Barrett, L. F., Lindquist, K. A., Bliss-Moreau, E., Duncan, S., Gendron, M., Mize, J., & Brennan, L. (2007). Of mice and men: Natural kinds of emotions in the mammalian brain? A response to Panksepp and Izard. *Perspectives on Psychological Science*, *2*(3), 297–312. (Publisher: SAGE Publications Sage CA: Los Angeles, CA)

Barrett, L. F., & Satpute, A. B. (2019). Historical pitfalls and new directions in the neuroscience of emotion. *Neuroscience letters*, *693*, 9–18. (Publisher: Elsevier)

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, *16*(7), 419–429. (Publisher: Nature Publishing Group)

Barrett, L. F., Wilson-Mendenhall, C. D., & Barsalou, L. W. (2015). The conceptual act theory: A roadmap.
(Publisher: Guilford Press)

Barros, P., Churamani, N., Lakomkin, E., Siqueira, H., Sutherland, A., & Wermter, S. (2018). The omg-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–7). IEEE.

Bhanoo, S. N. (2010, December). Humans, Like Animals, Behave Fearlessly Without the Amygdala. *The New York Times*.

Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. *arXiv preprint arXiv:1712.03586*.

Bohan, J. S. (1993). Regarding gender: Essentialism, constructionism, and feminist psychology.

Psychology of women quarterly, *17*(1), 5–21. (Publisher: SAGE Publications Sage CA: Los Angeles, CA)

Brescoll, V. L. (2016). Leading with their hearts? How gender stereotypes of emotion lead to biased evaluations of female leaders. *The Leadership Quarterly*, *27*(3), 415–428. (Publisher: Elsevier)

Cacioppo, J. T., & Gardner, W. L. (1999). Emotion. *Annual review of psychology*, *50*(1), 191–214. (Publisher: Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA)

Calvo, R. A., D'Mello, S., Gratch, J. M., & Kappas, A. (2015). *The Oxford handbook of affective computing*. Oxford University Press, USA.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900–E7909. (Publisher: National Acad Sciences)

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, *18*(1), 32–80.

DeLamater, J. D., & Hyde, J. S. (1998). Essentialism vs. social constructionism in the study of human sexuality. *Journal of sex research*, *35*(1), 10–18. (Publisher: Taylor & Francis)

Descartes, R. (1649). The Philosophical Works of Descartes. In *The Passions of the Soul*. Dover Publications.

Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, *18*(4), 407–422. (Publisher: Elsevier)

Dhall, A., Goecke, R., Joshi, J., Wagner, M., & Gedeon, T. (2013). Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI '13* (pp. 509–516). Sydney, Australia: ACM Press. doi:

10.1145/2522848.2531739

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2011, November). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (pp. 2106–2112). Barcelona, Spain: IEEE. doi: 10.1109/ICCVW.2011.6130508

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169–200. (Publisher: Taylor & Francis)

Ekman, P. (1993). Facial expression and emotion. *American psychologist*, *48*(4), 384. (Publisher: American Psychological Association)

Ekman, P. (2020). *Lie To Me \textbar Paul Ekman*. (Library Catalog: www.paulekman.com)

Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion review*, *3*(4), 364–370. (Publisher: Sage Publications Sage UK: London, England)

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, *17*(2), 124. (Publisher: American Psychological Association)

Ekman, P., & Friesen, W. V. (1978). *Facial action coding systems*. Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and an integration of findings* (Vol. 11). Elsevier.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). Facial action coding system: The manual on CD ROM. *A Human Face, Salt Lake City*, 77–254.

Ekman, P., & Keltner, D. (1997). Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, 27–46.

Ekman, P. E., & Davidson, R. J. (1994). *The nature of emotion: Fundamental questions*. Oxford University Press.

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

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Plunkett, K., & Parisi, D.

(1998). *Rethinking innateness: A connectionist perspective on development* (Vol. 10). MIT press.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007, December). The World of Emotions is not Two-Dimensional. *Psychological Science*, *18*(12), 1050–1057. doi: 10.1111/j.1467-9280.2007.02024.x

Fox, A. S., Lapate, R. C., Shackman, A. J., & Davidson, R. J. (2018). *The nature of emotion: Fundamental questions*. Oxford University Press.

Fragopanagos, N., & Taylor, J. G. (2005). Emotion recognition in human–computer interaction. *Neural Networks*, *18*(4), 389–405. (Publisher: Elsevier)

Gendron, M., Crivelli, C., & Barrett, L. F. (2018). Universality reconsidered: Diversity in making meaning of facial expressions. *Current directions in psychological science*, *27*(4), 211–219. (Publisher: Sage Publications Sage CA: Los Angeles, CA)

Gendron, M., Hoemann, K., Crittenden, A. N., Mangola, S. M., Ruark, G. A., & Barrett, L. F. (2020). Emotion perception in Hadza Hunter-Gatherers. *Scientific reports*, *10*(1), 1–17. (Publisher: Nature Publishing Group)

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion*, *14*(2), 251–262. doi: 10.1037/a0036052

Gifford, R., & O'Connor, B. (1987). The interpersonal circumplex as a behavior map. *Journal of Personality and Social Psychology*, *52*(5), 1019. (Publisher: American Psychological Association)

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological bulletin*, *117*(1), 21. (Publisher: American Psychological Association)

Gong, X., Huang, Y.-X., Wang, Y., & Luo, Y.-J. (2011). Revision of the Chinese facial affective picture system. *Chinese mental health journal*. (Publisher: Chinese Mental Health)

Gordon, S. L. (1991). 12 The socialization of children's emotions: Emotional culture, competence, and exposure. *Children's understanding of emotion*, 319. (Publisher: CUP Archive)

Griffiths, P. E. (2013). Current emotion research in philosophy. *Emotion Review*, *5*(2), 215–222. (Publisher: SAGE Publications Sage UK: London, England)

Gunes, H., Schuller, B., Pantic, M., & Cowie, R. (2011). Emotion representation, analysis and synthesis in continuous space: A survey. In *Face and Gesture 2011* (pp. 827–834). IEEE.

Harvey, P. (2012). *An introduction to Buddhism: Teachings, history and practices*. Cambridge University Press.

Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, *6*(2), 156–166. (Publisher: IEEE)

Hergenhahn, B. R., & Henley, T. (2013). *An introduction to the history of psychology*. Cengage Learning.

Hernandez, J., Hoque, M., Drevo, W., & Picard, R. W. (2012). Mood meter: Counting smiles in the wild. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 301–310).

Hester, N., & Gray, K. (2018). For Black men, being tall increases threat stereotyping and police stops. *Proceedings of the National Academy of Sciences*, *115*(11), 2711–2715. (Publisher: National Acad Sciences)

Hoemann, K., Crittenden, A. N., Msafiri, S., Liu, Q., Li, C., Roberson, D., ... Feldman Barrett, L. (2019). Context facilitates performance on a classic cross-cultural emotion perception task. *Emotion*, *19*(7), 1292. (Publisher: American Psychological Association)

Hoemann, K., Devlin, M., & Barrett, L. F. (2020). Comment: Emotions Are Abstract, Conceptual Categories That Are Learned by a Predicting Brain. *Emotion Review*, 1754073919897296. (Publisher: SAGE Publications Sage UK: London, England)

Hoemann, K., Khan, Z., Feldman, M., Nielson, C., Devlin, M., Dy, J., ... Quigley, K. (2020). Context-aware experience sampling reveals the scale of variation in affective experience. (Publisher: PsyArXiv)

Hoemann, K., Xu, F., & Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental psychology*, *55*(9),

1830. (Publisher: American Psychological Association)

Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science*, *2*(3), 260–280. (Publisher: SAGE Publications Sage CA: Los Angeles, CA)

Izard, C. E. (2009). Emotion theory and research: Highlights, unanswered questions, and emerging issues. *Annual review of psychology*, *60*, 1–25. (Publisher: Annual Reviews)

Kapoor, A., Qi, Y., & Picard, R. W. (2003). Fully automatic upper facial action recognition. In *2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443)* (pp. 195–202). IEEE.

Keltner, D., & Gross, J. J. (1999). Functional accounts of emotions. *Cognition & Emotion*, *13*(5), 467–480. (Publisher: Taylor & Francis)

Kervadec, C., Vielzeuf, V., Pateux, S., Lechervy, A., & Jurie, F. (2018). Cake: Compact and accurate k-dimensional representation of emotion. *arXiv preprint arXiv:1807.11215*.

Kleinberg, B., van der Vegt, I., & Mozes, M. (2020). Measuring emotions in the COVID-19 real world worry dataset. *arXiv preprint arXiv:2004.04225*.

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., & Pantic, M. (2017). AFEW-VA database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, *65*, 23–36. (Publisher: Elsevier)

Kosti, R., Alvarez, J., Recasens, A., & Lapedriza, A. (2019). Context Based Emotion Recognition using EMOTIC Dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. doi: 10.1109/TPAMI.2019.2916866

LeDoux, J. (2003). The emotional brain, fear, and the amygdala. *Cellular and molecular neurobiology*, *23*(4-5), 727–738. (Publisher: Springer)

LeDoux, J. (2012). Rethinking the emotional brain. *Neuron*, *73*(4), 653–676.

LeDoux, J. E. (2014). Coming to terms with fear. *Proceedings of the National Academy of Sciences*, *111*(8), 2871–2878. (Publisher: National Acad Sciences)

Levenson, R. W. (1994). Human emotion: A functional view. *The nature of emotion: Fundamental questions*, *1*, 123–126.

Levenson, R. W. (2011). Basic emotion questions. *Emotion review*, *3*(4), 379–386. (Publisher: Sage Publications Sage UK: London, England)

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives*, *21*(2), 153–174.

Li, B., Mehta, S., Aneja, D., Foster, C., Ventola, P., Shic, F., & Shapiro, L. (2019). A Facial Affect Analysis System for Autism Spectrum Disorder. In *2019 IEEE International Conference on Image Processing (ICIP)* (pp. 4549–4553). IEEE.

Lien, J. J., Kanade, T., Cohn, J. F., & Li, C.-C. (1998). Automated facial expression recognition based on FACS action units. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 390–395). IEEE.

Lindquist, K. A., Siegel, E. H., Quigley, K. S., & Barrett, L. F. (2013). The hundred-year emotion war: Are emotions natural kinds or psychological constructions? Comment on Lench, Flores, and Bench (2011).
(Publisher: American Psychological Association)

Lombardo, N. E. (2011). *The logic of desire: Aquinas on emotion*. CUA Press.

Manning, H. E. (1874). *Sin and Its Consequences*. Burns and Oates.

Mareschal, D., Johnson, M. H., Sirois, S., Thomas, M. S., Spratling, M., Spratling, M. W., & Westermann, G. (2007). *Neuroconstructivism: How the brain constructs cognition* (Vol. 1). Oxford University Press.

Marinoiu, E., Zanfir, M., Olaru, V., & Sminchisescu, C. (2018). 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2158–2167).

Martinez, D. L., Rudovic, O., & Picard, R. (2017). Personalized automatic estimation of self-reported pain intensity from facial expressions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 2318–2327). IEEE.

McClelland, J. L., Rumelhart, D. E., & Group, P. R. (1986). Parallel distributed processing. *Explorations in the Microstructure of Cognition*, *2*, 216–271. (Publisher: MIT Press Cambridge, Ma)

McDuff, D., Kaliouby, R., Senechal, T., Amr, M., Cohn, J., & Picard, R. (2013). Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 881–888).

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2012, January). The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing*, *3*(1), 5–17. doi: 10.1109/T-AFFC.2011.20

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019, January). AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, *10*(1), 18–31. (Comment: IEEE Transactions on Affective Computing, 2017) doi: 10.1109/TAFFC.2017.2740923

Nosakhare, E., & Picard, R. (2020). Toward Assessing and Recommending Combinations of Behaviors for Improving Health and Well-Being. *ACM Transactions on Computing for Healthcare*, *1*(1), 1–29. (Publisher: ACM New York, NY, USA)

O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135.

Panksepp, J., & Biven, L. (2012). *The archaeology of mind: Neuroevolutionary origins of human emotions (Norton series on interpersonal neurobiology)*. WW Norton & Company.

Panksepp, J., & Watt, D. (2011). What is basic about basic emotions? Lasting lessons from affective neuroscience. *Emotion review*, *3*(4), 387–396. (Publisher: Sage Publications Sage UK: London, England)

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 579–595.

Picard, R. W. (2000). *Affective computing*. MIT press.

Pinker, S. (2003a). *The blank slate: The modern denial of human nature*. Penguin.

Pinker, S. (2003b). *How the mind works*. Penguin UK.

Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, *89*(4), 344–350. (Publisher: JSTOR)

Plutchik, R. E., & Conte, H. R. (1997). *Circumplex models of personality and emotions*. American Psychological Association.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, *17*(3), 715–734. (Publisher: Cambridge University Press)

Rad, M. S., Martingano, A. J., & Ginges, J. (2018, November). Toward a psychology of *Homo sapiens* : Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405. doi: 10.1073/pnas.1721165115

Raskin, J. D. (2002). Constructivism in psychology: Personal construct psychology, radical constructivism, and social constructionism. *American communication journal*, *5*(3), 1–25.

Reisenzein, R., Studtmann, M., & Horstmann, G. (2013). Coherence between emotion and facial expression: Evidence from laboratory experiments. *Emotion Review*, *5*(1), 16–23. (Publisher: Sage Publications Sage UK: London, England)

Rudovic, O., Lee, J., Mascarell-Maricic, L., Schuller, B. W., & Picard, R. W. (2017). Measuring engagement in robot-assisted autism therapy: A cross-cultural study. *Frontiers in Robotics and AI*, *4*, 36. (Publisher: Frontiers)

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161. (Publisher: American Psychological Association)

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. doi: 10.1037/0033-295X.110.1.145

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of personality and social psychology*, *76*(5), 805. (Publisher: American Psychological Association)

Russell, J. A., Lewicka, M., & Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of personality and social psychology*, *57*(5), 848. (Publisher: American Psychological Association)

Russell, J. A., & Mehrabian, A. (1977, September). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11*(3), 273–294. doi: 10.1016/0092-6566(77)90037-X

Sales, P. P., & Lucena, R. (2000). DUELO: UNA PERSPECTIVA TRANSCULTURAL. MÁS ALLÁ DEL RITO: LA CONSTRUCCIÓN SOCIAL DEL SENTIMIENTO DE DOLOR. , 13.

Sayer, A. (1997). Essentialism, social constructionism, and beyond. *The Sociological Review*, *45*(3), 453–487. (Publisher: Wiley Online Library)

Scarantino, A. (2009). Core affect and natural affective kinds. *Philosophy of Science*, *76*(5), 940–957. (Publisher: The University of Chicago Press)

Scarantino, A. (2012). How to define emotions scientifically. *Emotion review*, *4*(4), 358–368. (Publisher: Sage Publications Sage UK: London, England)

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological review*, *61*(2), 81. (Publisher: American Psychological Association)

Sethu, V., Provost, E. M., Epps, J., Busso, C., Cummins, N., & Narayanan, S. (2019). The Ambiguous World of Emotion Representation. *arXiv preprint arXiv:1909.00360*.

Seyeditabari, A., Tabari, N., & Zadrozny, W. (2018). Emotion detection in text: A review. *arXiv preprint arXiv:1806.00674*.

Shields, S. A. (2013). Gender and emotion: What we think we know, what we need to know, and why it matters. *Psychology of Women Quarterly*, *37*(4), 423–435. (Publisher: Sage

Publications Sage CA: Los Angeles, CA)

Shields, S. A., & Shields, S. A. (2002). *Speaking from the heart: Gender and the social meaning of emotion*. Cambridge University Press.

Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., ... Yang, X. (2018). A review of emotion recognition using physiological signals. *Sensors*, *18*(7), 2074.

Siedlecka, E., & Denson, T. F. (2019). Experimental methods for inducing basic emotions: A qualitative review. *Emotion Review*, *11*(1), 87–97. (Publisher: SAGE Publications Sage UK: London, England)

Sorabji, R. (2000). *Emotion and peace of mind: From stoic agitation to Christian temptation*. Oxford University Press.

Stanislawski, K. (2019). The coping circumplex model: An integrative model of the structure of coping with stress. *Frontiers in psychology*, *10*. (Publisher: Frontiers Media SA)

Thanapattheerakul, T., Mao, K., Amoranto, J., & Chan, J. H. (2018). Emotion in a century: A review of emotion recognition. In *Proceedings of the 10th International Conference on Advances in Information Technology* (pp. 1–8).

Tovote, P., Fadok, J. P., & Lüthi, A. (2015). Neuronal circuits for fear and anxiety. *Nature Reviews Neuroscience*, *16*(6), 317–331. (Publisher: Nature Publishing Group)

Tracy, J. L., & Randles, D. (2011). Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emotion Review*, *3*(4), 397–405. (Publisher: Sage Publications Sage UK: London, England)

Trawalter, S., Todd, A. R., Baird, A. A., & Richeson, J. A. (2008). Attending to threat: Race-based patterns of selective attention. *Journal of Experimental Social Psychology*, *44*(5), 1322–1327. (Publisher: Elsevier)

Vielzeuf, V., Kervadec, C., Pateux, S., & Jurie, F. (2018). The Many Moods of Emotion. *arXiv preprint arXiv:1810.13197*.

Walcott, D. (1986). *Collected Poems, 1948-1984*. Macmillan.

Wang, Z., Lü, W., Zhang, H., & Surina, A. (2014). Free-labeling facial expressions and emotional situations in children aged 3–7 years: Developmental trajectory and a face inferior-

ity effect. *International Journal of Behavioral Development*, *38*(6), 487–498. (Publisher: Sage Publications Sage UK: London, England)

Wundt, W. M. (1907). *Outlines of psychology*. W. Engelmann.

Yik, M., Russell, J. A., & Steiger, J. H. (2011). A 12-point circumplex structure of core affect. *Emotion*, *11*(4), 705.

Zetlin, M. (2018, February). *Got a Poker Face? Employers are Using AI to Analyze Candidates' Facial Expressions and Personalities.* https://www.inc.com/minda-zetlin/ai-is-now-analyzing-candidates-facial-expressions-during-video-job-interviews.html.