The Ethics of Emotion in Artificial Intelligence Systems

Luke Stark
Faculty of Information and Media Studies
University of Western Ontario
London ON Canada
cstark23@uwo.ca

Jesse Hoey
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo ON Canada
jhoey@cs.uwaterloo.ca

ABSTRACT

In this paper, we develop a taxonomy of conceptual models and proxy data used for digital analysis of human emotional expression and outline how the combinations and permutations of these models and data impact their incorporation into artificial intelligence (AI) systems. We argue we should not take computer scientists at their word that the paradigms for human emotions they have developed internally and adapted from other disciplines can produce ground truth about human emotions; instead, we ask how different conceptualizations of what emotions are, and how they can be sensed, measured and transformed into data, shape the ethical and social implications of these AI systems.

CCS CONCEPTS

• Computing methodologies ~ Artificial intelligence ~ Philosophical/theoretical foundations of artificial intelligence • Social and professional topics ~ Professional topics ~ Computing profession ~ Codes of ethics • Human-centered computing ~ Human computer interaction (HCI) ~ HCI theory, concepts and models

KEYWORDS

emotion, affect, artificial intelligence, AI, machine learning, ML, ethics, norms, Basic Emotion Theory, Action Control Theory, affective computing, emotion AI, privacy, fairness, AI ethics

ACM Reference format:

Luke Stark and Jesse Hoey. 2021. The Ethics of Emotion in Artificial Intelligence Systems. In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442188.3445939

1 Introduction

Speculative and science fiction is replete with questions regarding the emotional lives of artificial beings. Yet contemporary machine learning-driven artificial intelligence (AI) systems have a much narrower view of human emotion than the complex questions posed in science fiction narratives. Computational analyses of psychological and behavioral data pertaining to human emotional expression have a surprisingly long history [31], an underappreciated diversity of methods [16, 108], and an increasingly critical role in social machine learning (ML) applications [28, 114]. AI/ML technologies are frequently used by social media platforms for modeling and predicting human emotional expression as signaling interpersonal interaction and personal preference [22]. In sectors including mental health care [20], personal health and wellness [28], education [124], hiring [129], automotive design [123], and national security [119] emotion detection and analysis is a rapidly growing sector for AI/ML systems [75].

While the fairness, accountability, and ethical and social impacts of ML/AI systems have become major topics of both public discussion and academic debate [8, 13, 18, 35, 60, 81], the ethical dimensions of AI/ML used to analyze human affective and emotional expression have been largely under-theorized in these conversations [3, 27, 44, 75, 114]. Given the increasing ubiquity of these systems, the ethics of affect/emotion recognition, and more broadly of so-called "digital phenotyping" [57] must play a larger role in current debates around the political, ethical and social dimensions of AI/ML.

Here we develop a taxonomy of the relevant conceptual models of human emotion and of proxy data for emotional expression; we then outline the ways the models of emotion and the proxy data collected according to these models influence design decisions made by the technologists creating AI/ML systems, and how these decisions raise broader questions about these technologies' social impacts. We do not take computer scientists at their word that the paradigms for human emotions they have developed internally and adapted from other fields should be taken naively ground truth: instead, we ask how different conceptualizations of human emotions shape the ways human values are built into and expressed by AI/ML systems.

2 Definitions and Theories of Emotion

Affect and emotion are concepts subject to intense debate across numerous academic fields [112] [9]. The shorthand term "emotion" is used to describe a compound phenomenon variously consisting of evaluative, physiological, phenomenological, expressive,

FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445939

behavioral, and mental components. Psychologist Jerome Kagan describes emotion as comprised of a variety of interrelated human phenomena: affect (or in Kagan's terms, "a change in brain activity to select incentives"); feeling or sensation ("a consciously detected change in feeling that has sensory qualities"); emotion proper ("cognitive processes that interpret and/or label the feeling with words"); and reaction ("a preparedness for, or display of, a behavioral response") [133]. Affect theorist Deborah Gould describes affects as "nonconscious and unnamed, but nonetheless registered, experiences of bodily energy and intensity that arise in response to stimuli" [42], and emotion as "what from the potential of [affective] bodily intensities gets actualized or concretized in the flow of living" [42]. A mood, another term often mobilized in the space of emotion tracking and quantification, is "a pervasive and sustained emotion that colors the perception of the world," [74], one that can pass quickly or stay for long periods of time.

This plethora of definitions points to what the philosopher of emotion Jesse Prinz describes as the dual concerns of the "problem of parts" and "the problem of plenty" in theories of emotion. The "problem of parts" describes the challenge of determining what component or components of emotion, be they evaluative, physiological, phenomenological, expressive, behavioral, or mental, are essential to its definition and detection in a particular context. The "problem of plenty" asks, if multiple components are essential to understanding emotions, how these various components hang together in practice [96]. As Prinz puts it, "the Problem of Parts asks for essential components, and the Problem of Plenty asks for an essential function of emotions in virtue of which they may have several essential components" [96]. Conceptual models of emotion tend to break down into different camps foregrounding one or another of emotion's "parts." These broad camps include understanding emotion as especially grounded in a) experienced feeling states, as b) evaluative signals connected to other forms of human perception of social cues, and c) as intrinsically motivating drives [104].

Affect and emotion are not the same phenomenon, yet in computer science the terms are at times treated interchangeably. This elision is always definitional (though sometimes inadvertent): human emotion is understood by computer scientists largely through the expression of biophysiological signals such as facial expression, gait, or blood conductivity [16, 75, 93]. Developers of AI systems have adapted this elision as a solution, if a potentially problematic one, to the Problems of Parts and Plenty. As we develop in Sections 4-6 below, the various conceptual models of emotion have incompatible ethical and social valences depending on which "parts" are emphasized in technical systems, and on how designers choose to reduce the polysemy of emotion for the convenience of technical constraint and business exigency.

2.1 Emotion as Motivating Drive

A robust tradition in affective science subscribes to the view that emotions are "distinctive motivational states" and "internal causes of behaviors aimed at satisfying a goal" [104] (s.8 p.1). In this paradigm, the human body is a palimpsest on which to read the eruptions of emotion. Scarantino and de Sousa argue that the motivational tradition is particularly concerned with understanding how emotions motivate human actions: in opposition to what he considered the lack of common sense in the James-Lange theory of emotions as derived from physiological phenomena, philosopher John Dewey argued in 1895 that emotions were not just experiences, but experiences with a purpose. The centrality of behaviorism in American psychology for much of the first half of the 20th century pushed most questions of motivation to the backburner [107]. However, the notion of emotions as motivating states reappeared in the 1950s and 60s with the work of American psychologist Silvan Tomkins [122] on affect, and the subsequent development of Basic Emotion Theory (BET) by Tomkins' students Paul Ekman [34] and Carrol E. Izard [55]. Tomkins argued that innate affects, or primarily physiological changes in brain and hormone activity, are hardwired into humans, identifying nine such basic affective programs: interest, enjoyment, surprise, fear, anger, distress, shame, contempt and disgust [104] (s.8 p.2).

Tomkins' students, in particular Ekman, carried these arguments further: in the 1970s, Ekman performed a series of comparative behavioral experiment which he claimed proved certain basic emotions were reliably recognizable in facial expressions in different populations around the world. As such, Ekman suggested, "there should be bodily signatures for each basic emotion consisting of highly correlated and emotion-specific changes at the level of facial expressions, autonomic changes and preset and learned actions" [104] (s.8 p.5). In other words, Basic Emotion Theory argues human emotional expression is universal, reliably legible, and critically, difficult to falsify—because emotions are understood as motivational drives deriving from biophysiological processes, they are difficult to conceal from the expert eye. The ethologist Alan Fridlund [39] has developed a notable alternative paradigm for understanding emotion as motivational, known as the Behavioral Ecology view: in this theory, all externalized forms of human emotional expression are better understood as social displays, always suggesting some sort of social motivation or function (e.g. a performative "suggestion") but which provide no evidence regarding the interior mental or motivational states of the expressor [68].

2.2 Emotion as Evaluative Signal

A second tradition, grounded largely in the cognitive revolution of the 1960s [125], treats emotions either as directly constitutive of cognitive states, or caused by them. Evaluative theories suggest that human emotions are one important category of evidence underpinning human thought and action, and that emotions are largely contingent on social contexts. These evaluative theories define emotions as primarily cognitive phenomena, and as "being (or involving) distinctive evaluations of the eliciting circumstances" (s.2 p.4). Central to this view is that emotions have "intentionality": that humans direct our emotions at particular objects (a view synthesized and popularized in the 1870s by the German philosopher Franz Brentano). Whether understood

philosophically as either simple judgments or complex appraisals, the evaluative tradition understands emotion as in some way connected to human judgment (though there is further broad debate as to of what "judgment" itself consists). In psychology and affective science, Scarantino and de Sousa, reviewing [6] and [66], connect the evaluative tradition to the rise of appraisal theory. Appraisal theory is concerned with developing "accounts of the structure of the processes that extract significance from stimuli and differentiate emotions from one another" [104] (s.6 p.5), though such differentiation does not conflict with understanding emotions as evaluative, experiential, or motivational *per se*.

2.3 Emotion as Felt Experience

A third tradition in the philosophy of emotion is to treat emotions as primarily experiential: as Scarantino and de Sousa put it, this "Feeling Tradition takes the way emotions feel to be their most essential characteristic and defines emotions as distinctive conscious experiences" [104] (s.2 p.4). In essence, this way of understanding emotion likens it to other felt experiences such as taste, pain, or other embodied sensations. "We feel," wrote the nineteenth-century philosopher and psychologist William James, "sorry because we cry, angry because we strike, afraid because we tremble, and not that we cry, strike, or tremble, because we are sorry, angry, or fearful" [58]. This view broadly typified philosophical theories of the passions from Aristotle through to the nineteenth century, contrasting emotions with the discipline typifying logical reasoning and reflection. In 1884, James proposed a variation of this longstanding view, positing that emotions were a specific kind of subjective experience [58]: under what became known as the James-Lange hypothesis, emotions were, "sensory feelings constituted by perceptions of changes in physiological conditions relating to the autonomic and motor functions" (s.3 p.3). In this view, perception of emotions as such are derived from physiological process (exemplified by James' famous invocation that we are afraid of a bear because we are impelled by reflex to run away from it).

2.4 Hybrid Theories

Basic Emotion Theory has been enormously influential since its promulgation by Ekman in the 1970s, not least in recent work on artificial intelligence, robotics, and computer science more broadly. However, BET has also come under sustained critique from a number of quarters, including from proponents of what might be best termed "hybrid" theories of emotion mixing elements of the experiential, evaluative, and motivational traditions. Both James Russell [102, 103] and Lisa Feldman Barrett [9, 10] have vigorously critiqued Basic Emotion Theory, arguing that while core affect is an important component of emotional experiences, emotion itself emerges out of human evaluative and experiential assessments of affective states in particular social contexts [95].

Scarantino and de Sousa note that philosophical and psychological theories understanding emotion as a form of evaluative assessment and as felt experience are increasingly intertwined in "hybrid" approaches, with "the former now identifying emotions as evaluative perceptions with a distinctive phenomenology and the latter identifying emotions as evaluative feelings with a distinctive intentionality" [104] (s.7 p.2). Prinz's [96] perceptual theory of emotion is one such hybrid approach, treating emotions as evaluative perceptions grounded both in physiological affective changes and in the particular social context of the object or person being perceived. Sociologist Arlie Russell Hochschild, who developed the concept of "emotional labor," likewise presents a hybrid theory articulating emotions as a mix of physiological and social signals: in her words, "every emotion has a signal function" [50]. Affect Control Theory (ACT) [49, 110, 111] a form of structural symbolic interactionism mapping so-called patterns of emotional salience (s.7.3, p.1)—or typical social evaluations of affective response of actors and behaviors—is a third hybrid theory incorporating elements of evaluative and experiential models. [70]. A Bayesian extension, BayesACT [108] combines this conceptual framework with a motivational model based in decision theory.

3 Proxy Data for Emotional Expression

If the breadth of theories for understanding what emotions are is not daunting enough, the empirical evidence that outside observes use to identify and understand emotion also varies widely. This evidence is often quantitative data. In the context of digital and artificial intelligence system, practitioners have used or suggested various types of proxy data for analysis of emotional expressions [63]; the decision to collect a particular proxy often depends on what conceptual model of emotion the designers of the underlying agent or system have adopted [113, 114]. These various proxies for emotional expression include physiological data [91], such as facial expression [47], gait, or infrared emanations and haptic and proprioceptive data (such as skin conductivity, blood flow, and body velocity) [86]; audio data (such as the vocal tone and cadence) [99] [20]; behavioral data collected over time [57]; and semantic signifiers of emotional expression, (including written words and graphic means such as emoji and emoticons) [2].

Historian of medicine Otniel Dror has expertly documented the origins of what he terms "emotion as number" in the late nineteenth century, through the fusion of empirical physiology, a homosocial culture of masculine expertise, and a desire to contrast medical studies of emotion with the mobilization of feeling by anti-vivisectionists [32]. The full genealogy of how this focus on physiological signs became understood as standing in for—and in some accounts, consisting of—human emotion in computer science is largely outside the scope of this paper [31, 33, 36, 73, 117, 125], but is relevant as an example of how contingent definitions of one aspect of emotional response, in this case, bodily changes, can become a dominant paradigm for explicating emotions through its utility to particular sets of experts (in this case, physiologists, cognitive scientists, and their computationally inclined descendants).

Depending on the conceptual model of emotion at hand, the accuracy and suitability of a particular type of data taken as proxy

for interior emotional states will—as with any type of social data—vary widely [84]. According to some of the theoretical traditions described above, some forms of proxy data, such as heart rate or facial expression, can say little about a person's emotional state—while for others, these forms of data are key indicators of such. The current paradigm around "affective computing" was sparked largely by the work of MIT's Rosalind Picard, whose [91] book *Affective Computing* argued for emotion as a topic worthy of examination by computer scientists. Picard's approach focused on, as noted, physiological signals such as heart rate and blood flow as proxies for emotions and sought to translate these signals into data a computer could find legible and tractable for analysis—treating human emotional expression as another form of digitizable information. Scholars in social computing and critical HCI contested this approach almost immediately, noting such an "informational reading… systematically ignores a second set of concerns which focus on emotion as it is interactionally and culturally constituted" [16]. This interactional approach to the digital mediation of emotional expression and interaction, laid out in a series of papers by critical HCI scholars including Kirsten Boehner, Phoebe Sengers, Katherine Isbister and Kia Höök among others [15-17, 52, 54, 67], emphasizes the centrality of supporting human emotive interaction through a diverse array of digital technologies, instead of focusing narrowly on the sensing, tracking, quantification and analysis of those interactions via computational data. Central to the interactional approach is the variability and dynamism of human social and emotional relationships: past performance being no guarantee of future results in life, it should not, as a design criterion, be emphasized when dealing with the digitally traced "emanations" [62] of our social and emotional lives (for a further summary, see [30]).

4 A Taxonomy of Theories and Proxy Data

We taxonomize AI systems and products for tracking, interpreting, and modeling human emotional expression (**Table 1**) along two axes: a) the conceptual models of emotion on which these systems are grounded, either explicitly or implicitly; and b) the types of data these systems collect and use as proxies to assess human emotional states. For simplicity and given their increasing interconnection, we combine the Feeling/Evaluative models in the table below.

4.1 Motivational Theories of Emotion and AI

Motivational theories of emotion, in particular Basic Emotion Theory, have proven particularly amenable to adaptation into AI systems. These systems are grounded in the notion that humans have regular, universal, and traceable emotional expressions and reactions, legible across a wide range of proxy data; they make up the lion's share of "Emotional AI" technologies currently available.

A large segment of the discipline of "affective computing," involving tracking and analyzing a variety of bio-signals like heart rate, is grounded in Basic Emotion Theory [26, 91, 92, 94]. Models powering facial analysis technologies are designed around Ekman's thesis regarding the legibility of basic emotions through facial expression [40], and analyze large databases of human faces [21], using various AI techniques to estimate emotional expression (along with other characteristics like age and gender) [105]. These systems are increasingly commercially available in areas such as human resource management, advertising [76] and education [12]. Lisa Feldman Barrett has been a strong critic of the application of Basic Emotion Theory, both as a scientific consensus [68], and particularly in its application in digital systems [11]. In a major recent review, Feldman Barrett and co-authors critiqued the proliferation of emotion detection in facial recognition technologies (FRTs) by calling the underlying generalizability and robustness of Basic Emotion Theory's assumptions into question [11]. The authors declared that "When facial movements do express emotional states, they are considerably more variable and dependent on context than the common view allows" [11]. Despite this caution, however, emotion recognition in FRT continues to generate strong commercial interest.

Proxy data for emotional expression from other sources, such as recorded audio of the human voice, or recordings of the electrical conductivity of the skin [94], are also often analyzed under assumptions grounded in the Motivational paradigm. The proliferation of such digital data, coupled with the popularity of BET as a computationally tractable theory among Silicon Valley startups, has begun to underpin various assumptions about a wide range of digitally tracked behaviors and their purported connections to human motivation. In the last five years, the broader universe of such analysis for all sorts of behavioral and social data, including around emotion, are increasingly classed under the term "digital phenotyping," or "measuring behavior from smart phone sensors, keyboard interaction, and various features of voice and speech" [53]. The term itself was coined as a term in a 2015 paper by Sachin Jain and fellow physicians at the Harvard Medical School, based on a concept drawn from the 1982 book *The Extended Phenotype* by evolutionary biologist Richard Dawkins. Dawkins had argued for the extension of the notion of phenotype from the set of observable characteristics of an individual "to include all effects that a gene has on its environment inside or outside of the body of

Table 1.	***Types of Data Collected***	**Physiological**	**Auditory**	**Haptic**	**Behavioral**	**Semantic**	**Social**
Model of Emotion							
Motivational		Facial recognition [47]	Vocal-diagnostic tech [99]	Galvanic response [86]	Digital phenotyping [53]	Sentiment analysis [2]	
Experiential/ Evaluative		Brain/machine Interfaces [48]			E-A ML [119]	CBDTE [98]	BayesACT [108]

the individual organism" [57]. Jain and coauthors reinterpreted Dawkins' use of the term "phenotype" loosely, to refer to any

manifestation or emanation traceable by a digital sensor. As such, large-scale computational analyses now implicitly equate patterns of behavior with intrinsic interior states, including motivation; under the Motivational paradigm, these analytics can in theory correlate a wide array of data produced by humans with inferred emotional states [97].

It is also possible for the developers of AI systems to treat proxy data around emotion as predominantly informing outside observers about social displays of emotion (as per Fridlund's Behavioral Ecology view); observers need not make any assumptions about internal affective states to develop an externally consistent view of emotional interactions. Sentiment analysis of textual or graphic representations of emotion, such as dictionaries of emotion terms or emoji characters [127], present a case in point. Contemporary deep learning AI systems often analyze structured natural language data, and in the best case can develop sophisticated models allowing them to uncover patterns of emotive data in language. These systems have no capacity to understand the conceptual "motivation" behind a text; however, AI systems may not need to "understand" motivation if they have a sufficiently clear map of interactions. Facebook's introduction of "Reactions" icons in 2016, enabling users to "react" to all posts with one of six basic emotive symbols (themselves based on Ekman's BET), was aimed at developing such a network graph of user content irrespective of the motivation users had for reacting in the first place [114].

4.2 Experiential/Evaluative Theories of Emotion and AI

Conceptualizations of emotion grounded in the experiential/evaluative tradition are less common in current AI systems, but those that do exist provide notable and instructive contrasts to those of the dominant Motivational paradigm. One method grounded in the evaluative tradition for modeling human social and emotional interactions computationally is Bayesian Affect Control Theory or BayesACT [108]. Affect Control Theory, initially developed by social psychologist David R. Heise [49], is a quantitative sociological method akin to structural symbolic interactionism [70]. In an ACT analysis, the interactions between actors, behaviors, and settings are mapped by an observer: in doing so according to ACT, "the observer realizes if the situation is aligned with cultural norms or represents a deflection from cultural norms based on the affective sentiments. In the case of deflection, the observer tries to restore a coherent definition of the situation" [110]. ACT implicitly defines a morality concept which is embedded in these cultural norms [72]. Developed out of collaborations between sociologists and computer scientists [108], BayesACT combines ACT with Bayesian probabilistic decision theory to make evaluative predictions about what emotions are appropriate for a variety of typical situations. These predictions enable virtual agents to better calibrate their responses to users, letting the system detect emotional cues and respond appropriately. BayesACT further provides a theory of action motivation as it couples connotative meanings (cognitive appraisals of sentiments) with decision theoretic reasoning over denotative states. Sentiments are used to guide action towards socially normative behaviors. These techniques are being tested in areas such as cognitive assistive technologies for persons with dementia that are functionally and emotionally aligned with their target users [64], and facilitator agents in social networks aimed at promoting efficient and inclusive group processes.

Other scholars have proposed simulating the appraisal or evaluative aspects of emotion as elements of algorithmic learning models themselves. For instance, the Computational Belief and Desire Theory of Emotions (CBDTE) developed by Reisenzein [98] argues that emotions are caused by a combination of cognitive evaluations (beliefs) and conative motivations (desires), and that such a theory can be modeled formally within learning algorithms used to analyze natural language. In a similar vein, Emotion-Augmented Machine Learning (E-A ML) focuses on developing computer models that incorporate simulations of emotion concepts into the machine learning process [119]. Such techniques seek to enhance the performance of reinforcement learning systems by constraining their evaluative behavior using such simulated emotion concepts as anxiety. In theory, such models would perform more like humans, using emotions like anxiety as a heuristic filter to focus on the best available course of action. Finally, recent work on Brain-Machine Interfaces [69] has suggested the possibility of translating the brain patterns of emotional experience into digital data, either through direct implants or mechanisms such as audio waves [48].

5 Emotion in Current AI Ethics Debates

The lack of consensus around both conceptual models and empirical proxies for emotion has important normative implications for the AI systems reliant on them. The implied ethical and social responsibilities for human persons vary based on the causal models of emotion at issue. The social and ethical weights given to subjective human experience, motivation, or belief are shaped by how an observer understands the potential causes for emotional expression, the evidence of such expression, and how emotion as a normative force is accounted for when considering the impacts of values such as fairness, transparency, or accountability in AI systems.

Andrew McStay [7, 75, 76, 78] has been among the most active voices at the intersection of scholarship on emotion and AI ethics. McStay's work points to the proliferation of systems for tracking, inferring, and measuring signals of human emotive expression, what McStay terms "Emotional Artificial Intelligence (EAI)," as necessitating a broader focus on both the technical affordances and social impacts of these technologies. With Pamila Pavliscak, McStay has developed a set of guidelines for the ethical use of EAI technologies, designed explicitly to enable companies to "innovate ethically as well as legally" in the area [77]. In line with the wider

proliferation of AI ethic codes, statements of principle and similar envisioning documents [38, 43, 59], the EAI guidelines present a high-level thematic checklist for anyone engaged with "data about human emotion." Noting that, "Emotional artificial intelligence has significant personal, interpersonal, and societal implications," the authors' varied instructions for practitioners are divided by putative relevance to the individual person, to relationships, and to society. McStay and Pavliscak's guidelines also include a number of salutary suggestions for "taking action" as a practitioner against unethical design decisions, a rarity in the ethics guidelines genre (including suggesting either saying no joining to such efforts or leaving an institution based on ethical scruples).

A number of the document's prescriptions, including all those under the listing of the "Personal" implications for EAI, are important but broadly applicable to all identifying data. Under the heading of implications for "Relationships," however, the guidelines list three checklist items specifically focused on emotion (as well as one guideline aiming to safeguard trust in non-human actors, and one dealing with procedures for identifying mental health challenges in users). These three items are: "Understands that physical display of emotion is only one facet of emotion"; "Recognizes that past expression doesn't always predict future feeling"; and "Considers stereotypes and assumptions about emotion that materially affect a person or group." Under "Societal" implications, the guidelines provide two items focused on emotion: "Recognizes the lack of globally objective agreement on emotion," and "Recognizes that collecting data about emotion in public spaces may be unwanted or invasive." These five items (one-third of the total number in the guidelines) deal explicitly with emotion, as opposed to the wider universe of concerns around the use of digital data for tracking, profiling, and behavioral nudging in the digital economy [4, 24, 114]. These details of McStay and Pavliscak's schema are worth examining closely, in how the guidelines implicitly define emotion, for what they foreground as ethical and social challenges unique to EAI as opposed to AI more generally, and what they omit—all elements indexical to the broader gaps in conversations around emotion and ethics in AI.

McStay and Pavliscak's call for practitioners to reflect on the fact that "physical display of emotion is only one facet of emotion" is sound, and poses practical challenges for practitioners steeped in the particular, and narrow, definitional discourses around Basic Emotion Theory common in computer science. In a similar vein, the EAI guideline item prompting practitioners to recognize that, "past expression doesn't always predict future feeling" refers implicitly to larger debates in computer science regarding how digital media technologies should understand and approach the longitudinal stability human of emotional expression as an element in technical design. The guidelines also ask if practitioners have considered "stereotypes and assumptions about emotion that materially affect a person or group." Across these three questions, the guidelines assume a stable and common definition or understanding of what emotion is in the first place.

The EAI guidelines suggest further that, "collecting data about emotion in public spaces may be unwanted or invasive." This guidance points to the particular sensitivity many people ascribe to emotion and emotional expression [3], but also elides the fact that certain emotional expression and the boundaries of public and private space vary and intersect in granular and sometimes surprising ways [118]. For instance, the oft-referenced 2014 Facebook emotional contagion study [65], which came under fire in the media for manipulating the semantic emotive content of user news feeds, was criticized precisely because Facebook users had a different felt understanding of the privacy of their interactions than did the researchers and the platform [113]. As Siva Vaidhyanathan [132], Frank Pasquale [87], Tero Karppi [61], and others have noted, Facebook's business model entails making, and shaping, assumptions about human emotion precisely so it can affect the choices of both individuals and groups on the platform to engage and interact. This molding of what Zizi Papacharrisi terms "affective publics" [85] does not always have predictable results and is always subject to counter-pressures from individuals and groups themselves [46], but even under an interactional model of emotion is always a factor in design and deployment. Moreover, the varying cultural contexts of emotional norms, variations, and interactions makes broad assumptions about emotional universalism not just unwise, but actively deleterious.

Finally, the guidelines ask practitioners to recognize, "the lack of globally objective agreement on emotion." This item points to the central problem at hand, both for AI practitioners building systems engaging with data about human emotion and perhaps for McStay and Pavliscak's guidelines themselves: the wide divergence of opinion in both philosophy and in the sciences regarding what emotions are, means recognizing that diversity of opinion in the abstract is insufficient without considering how those differences might implicate AI design and deployment decisions, with their attendant ethical valences and social effects.

6 Analysis and Discussion

Digital data on human emotional expression not only imperfectly reflect the complexity of human emotional response; the conceptual models used to make sense of emotions themselves also imperfectly reflect that complexity, as per the Problem of Parts and the Problem of Plenty described by Prinz [96]. Both models and data are schematically representative of the multiple elements of human emotional experience, denoted imperfectly and quantified partially. As such, there will always be a gap between the model used and the experience lived, a problem increasingly well-articulated in critical studies of AI systems more broadly. As Selbst et al. observe, "abstracting away the social context in which these systems will be deployed [means] researchers miss the broader context, including information necessary to create fairer outcomes, or even to understand fairness as a concept" [111]. The particular complexity and personal sensitivity of human emotions makes this challenge especially salient, and potentially troubling, for emotion AI systems. While some of these challenges map to broader

concerns around biases and aporias in other forms of personal data, in particular healthcare data used in AI/ML analysis [41], the collection of data around emotional expression and emotion modeling in AI also present a number of unique additional normative challenges. These challenges include often implicit associations between human emotions and normative categories; the particular implied norms of certain emotion models such as Basic Emotion Theory; human emotions as a locus for online experimentation; the dangers of reifying particular emotion metrics; and the lack of scientific consensus underpinning the models used in AI systems intended to measure and model emotion.

6.1 The Opaque Normative Weight of Emotion

The most basic normative concern around the collection of data on human emotional expression and the computational analysis of that data stems from the ways in which emotions are associated, implicitly and explicitly, with human agency in theories of ethical decision-making. In other words, what normative weight do the various conceptual models we have already described place on emotions in light of their assumed relationship to human action? If emotion and other intuitive processes are understood to play a central role in ethical and moral judgments, then incorporating any proxy data for human emotion into an AI system takes on fraught normative importance. The choice of conceptual priors and the type of emotion data collected—indeed, the decision to design and deploy an AI system engaged with human emotion in any way at all—will invariably import particular norms and values into a technical system, ones that will affect the impacts of these systems in ways often unanticipated by designers and others responsible for their deployment.

Scarantino and de Sousa observe that emotions are often understood as impediments to rationality, and by extension to considered rational judgment: "Emotions," the authors write, "have long been thought to score poorly in terms of both cognitive and strategic rationality," the former "consisting of their ability to represent the world as it is," and the latter "consisting of their ability to lead to actions that promote the agent's interests" [104] (s.10.1). However, more recent scholarship in both psychology and neuroscience has highlighted the centrality and necessity of emotion as a component in "rational" cognition. Emotions "determine salience among potential objects of attention," and while this phenomenon has the potential to misdirect attention, it can also help sustain long-term planning and goal setting [9]. Self-reflexivity around emotional responses is central to their balanced contribution to "rational" outcomes, but such reflexivity also requires weighing "rational" or cognitive considerations against the signal function provided by the emotion states themselves. In recent work in affect control theory, the tradeoffs between rational and emotional cognition are considered primarily in light of the relative uncertainties between the two forms of mental process in emotionally charged situations; the ways in which the tradeoff is made may be culturally or individually dependent [51].

It thus matters both in terms of what investigators are measuring as a proxy for emotional or emotive expression, but also what investigators believe the responses measured mean about the interiority, judgments, and potential future actions of human beings. For the creators of emotion AI systems, the need to understand the subtleties around judging the normative significance of emotion adds a third layer of complexity on top of the two layers already described (what conceptual model is being used to explain what emotions are, and what data is being collected to determine how emotions are expressed). In practice, these three categories are interrelated: normative judgments can emerge from conceptual assumptions, themselves grounded in a particular interpretation of empirical data or the choice of what data is serving as proxy for emotive expression.

Emotions are not only irreducible to any one form of proxy data but are also subjective phenomena in part illegible to outside observers. As such, any AI system engaging with models or data about human emotions should be flagged immediately by oversight authorities as requiring heightened scrutiny around its social impacts and normative effects, irrespective of the context of use. Abstracting away the social context of an emotional expression presents a fundamental barrier to the comprehensive understanding of emotion; many current AI-based efforts to do so contribute to scientific overreach, unethical and anti-democratic experimentation and manipulation, and the internalization and reification by individuals of the same problematic metrics. The notion of predicting the individual emotional states of particular people using these systems is therefore always suspect, as even computerized iterations of evaluative models like Affect Control Theory will extrapolate typical emotion reactions which will not hold in all cases. Designers and developers should think twice before embarking on emotion AI projects: a necessary though not sufficient condition for such projects is clear alignment between conceptual models, data, norms, and aims.

6.2 The Troubling Norms of Basic Emotion Theory

It is also worth considering how the broad conceptual split between motivational theories of emotion and experiential/evaluative theories might shape how emotions are understood normatively, and the implications for AI/ML systems that incorporate one or the other of these conceptual models for emotion. The most obvious division comes around the motivational tradition's focus on emotions as causal phenomena. At least in Basic Emotion Theory, this view of emotions as motivational leads to an understanding of exteriorized emotional expression as "true" manifestations of inner emotional states, and as uncontrollable, and thus unfalsifiable symptoms of internal subjective impulses.

Given the popularity of BET as the conceptual underpinning for the design of AI/ML systems that track and categorize emotional expression, this fundamental division in the understanding of emotion's relationship to agency has an outsized effect on the ethical and social impacts of these AI systems as they are deployed

in practice. In the case of BET, this influence is unfortunate. A number of commentators [14] [27, 89, 116], one of us included [115], have argued that AI/ML systems used to analyze human faces, bodies, and gaits are engaged in a digitally-mediated form of physiognomy, the discredited nineteenth- and early twentieth-century practice of using people's outer appearance to infer inner character [29, 106]. Motivational theories of emotion align, however loosely, with both the legacy of physiognomy and the logics of contemporary facial recognition technologies (FRTs) and other similar systems [40].

Moreover, as critical scholars of both race and gender have argued, emotional expression is a key vector through which racist hierarchies and misogynist tropes are produced (or "discovered") routinized, and enforced [82, 126], often through the mobilization of motivational theories that purport to reveal "inferior" interiority through externalized emotive signals [101]. For instance, Kyla Schuller [109] articulates how discourses around the "biopolitics of feeling" developed in the nineteenth century equated emotional impressibility with civilizational refinement—and by extension, defined "primitive" subjects as ones whose emotion were both legible and predictable. Likewise, Otniel Dror's account of nineteenth century "emotion as number" points to how such constructed hierarchies were quantified and solidified through technical language [32, 33] .

While many of today's proponents of Basic Emotion Theory may not be aware of these historical genealogies, they are nonetheless impossible to discount, to say nothing of more recent critiques of the field [9, 68]. Basic Emotion Theory does not lend sufficient scientific evidence for even nuanced arguments grounded in the Motivational tradition, much less the often sensationalist claims made by many FRT providers that such systems are able to easily determine what individuals are "really feeling" [11]. BET's influence as the chief paradigm for the digital mediation and interpretation of human emotion is thus a major normative challenge for designers and regulators, and one which deserves heightened scrutiny from ethical and regulatory perspectives.

6.3 Emotion as a Locus of Experiment

A focus on emotional expression as a component of AI/ML analysis demonstrates the broader tendency of AI/ML researchers in corporate settings to perform *de facto* human subject research without attendant awareness of, or attention to the ethical complexities of such experiments [20]. In the 2020 documentary *The Social Dilemma*, for instance, much of the film's central message focuses on the ability of social media platforms to control users, and to sell this control to advertisers. Personal data is at the heart of the business models of many digital technology companies, and the collection of information about every aspect of a user's (or even a non-user's) interactions with a site or app is now well understood as a major privacy and civil society problem [83, 88, 128]. Data on human emotional expression is what Nicholas Terry terms "medically-inflected" data [120], possessing a sensitivity on par with data protected under regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, but rarely treated as such. Though outside the scope of this paper, current conversations on how to adapt research ethics protocols to digital contexts in both the academy and industry are only the start of a larger set of normative concerns around these data practices [37, 131].

Experimentation on the part of social media companies often entails modifying elements of their interfaces to affect users: these modifications can have many effects, including an increase in political polarization [19], subversion of existing consent regimes [131], and distrust of the modifications even if they are innocuous [79]. The popular furor around Facebook's 2014 "emotional contagion" study [65], and subsequent 2018 Cambridge Analytica scandal [100, 130], are exemplary of how behavioral experimentation around emotion online can have long-standing cultural consequences [114], they are also indexical to broader questions around the ethics of experimentation on the part of digital organizations on the emotional, psychological, and behavioral data of their users.

However, it is also important to note that such manipulations do not imply that these companies can build an accurate model of a person based on their affective responses, nor that they necessarily understand the causal relationships between the responder and the modified elements of their interface [56]. While one may deduce that an affective response arises from a change made on a social media platform, it is incorrect to claim this means one can infer or induce a person's "phenotype" from their affective responses. While these responses may have resulted, in part, from their reactions to the change made, it may also depend on many other factors that are context or person dependent. Since a *response* is a causal effect of the *change* and the *person*, experimenters can deduce that the *response* arose because of the *change* and can even measure the results of an A/B test to see which change is more apt at delivering a response —but cannot infer much about the person given a particular response to a change because of the unknown other factors impacting the affective response.

6.4 Reification and Interiorization of Models/Metrics

The conceptual models and proxy data for emotion in an AI system are not solely a concern for the system's designers. These models and data possess a descriptive power that is also a prescriptive one [71, 114]. Individuals often adjust their own attitudes to conform to an "objective" measure, in this case of emotional expression, that is in fact partial, constructed and potentially detached from lived experience. As such, the digital remediation of emotional expression has the potential to shift subjective normative frameworks for decision-making towards the emotional models, and implicit values, of technology firms, not of individuals as users and citizens.

In a recent qualitative study [3], Nazanin Andalibi and Justin Buss provide evidence for the outsized impacts the digital remediation of emotional expression can have on individuals. Andalibi and

Buss's respondents were highly aware of, and concerned with the potential power of these systems: "The majority of participants," the authors observe, "were uncomfortable with emotion recognition, and this discomfort was often related to concerns over privacy, consent, agency, and potential harm" [3]. Participants also pointed to a lack of accountability on the part of the designers and developers emotion recognition and analysis systems to engage with these embodied concerns. As Andalibi and Buss point out, there is not only a lack of recognition on the part of technologists regarding the multiple potential sources of data about emotion, and conceptual frameworks under which it is collected; there is also a lack of recognition of the diversity of human attitudes towards their own emotions, and to emotions as social phenomena impacted by digital remediation. Vernacular media products such as the 2015 Disney/Pixar film *Inside Out*, which depicts the brain and emotions through a combination of BET theory and metaphors of mediation [45, 121], are a further mechanism reinforcing the discursive power of particular emotion models in everyday discourse.

Individuals whose emotion data are being tracked and aggregated are thus often in a bind regarding how to respond to the analytic outputs of AI systems. If relationships between different variables correlate in the aggregate, there is a danger that modelers will assume the same relationship will also correlate at an individual level, an error known as the 'ecological fallacy' [90]. Tools like computational sentiment analysis are increasingly deployed to chart the "ambient sentiment" of groups such as Twitter users [5]. Yet a focus on aggregate modeling effaces the ways that individuals must modulate their emotional responses over time to conform to norms produced by the aggregating institutions, such as social media platforms: aggregated categories are represented back to individual users as norms against which they should perform. [114]. Individuals must preemptively position themselves as fluent in the emotional expressions, behaviors and gestures aligned with a platform's models, able to both conform to these classificatory schemes while caught in the everyday pressure to perform emotional expressions in non-virtual social situations.

The questions of aggregation and reification also cuts across global cultures, where the lack of "global" agreement on emotion in more ways than one also threatens to produce platform-driven regimes of emotive conformity. Anthropologists have long observed the cultural specificity of emotion norms and discourses [1], and the dangers of attempting to presume stability across heterogenous views of emotional expression [50]. While recent scholarly work has begun to document cultural variation in the use patterns for digital formats for emotive expression such as emoji [25], there is little comparative research on how the emotional models built into AI systems vary in their performance and interpretation based on particular geographical locations and cultural norms – let alone whether the use of these emotional models has begun to reflexively change those norms and homogenize emotional expression around the world. A number of scholars have examined the ways individuals perceive, understand, and interpret the workings of algorithmic systems [22, 23]. More research is needed to examine how the models of emotion most prevalent in representations of AI systems and digital media more broadly, such as BET, are affecting individual and collective subjective assessments of emotional agency, and how these changing subjective mental paradigms are shaping actions and behavior in diverse cultures worldwide.

6.5 Lack of Scientific and Normative Consensus as Disqualifying

Finally, attempts to quantify and standardize measures of emotion through its expression illustrates the wider conceptual difficulty in constituting ethical or normative guidelines around AI/ML systems shared broadly across communities and societies, due in part to what philosopher Thomas Nagel terms "the fragmentation of value" [80]. What components of emotion are most salient in a particular context? The lack of consensus raises the question of whether it is ever ethically appropriate to develop and deploy such systems for public consumption: if, for instance, the science of Basic Emotion Theory cannot support the claims its AI/ML proponents make, their incorporation into AI systems is potentially fatally flawed regardless of other ethical safeguards. This problem is analogous to broader debates around appropriate data collection versus appropriate data use in the digital privacy arena. While safeguards on the appropriate use of emotion recognition in AI systems are necessary, they are not sufficient, and a wider conversation around the deployment of emotion recognition systems in AI is vital given the potentially toxic social effects such technologies can produce [115].

7 Conclusion

The analytics of emotional expression highlight human emotion's centrality not just to ethical AI/ML systems, but also to these system's broader mediating effects on social and political community and cohesion through their everyday use. Human emotion is a complex topic, and analysis of its effects and impacts in AI/ML benefit from interdisciplinary collaboration. There is thus a critical and urgent need for scholars, policymakers, and technologists to understand the complexity of human emotions and the digital economy being built on them when designing, critiquing, and regulating AI systems. As research and commercial interest in "artificial emotional intelligence" (AEI) intensifies, we argue the particularities of how these systems are designed—including the models of emotion designers use to ground their models, and the types of proxy data for emotion they collect—matter greatly for the ethical appropriateness of such systems, and even whether they should be developed and deployed at all.

ACKNOWLEDGMENTS

Our thanks to Ben Green, Deborah Raji, Varoon Mathur, Casey Gollan, Alejandro Calcaño Bertorelli, Sarah Myers West, Erin McElroy, Elizabeth Kaziunas, Osonde Osoba, Fernando Diaz, Benjamin Fish, Asia Biega, Alexandra Olteanu, Nazanin Andalibi, Emily McBain-Ashfield, Jason Millar, Florian Martin-Bariteau,

and the other organizers of the WeRobot 2020 virtual conference for contributing to earlier versions of this piece; and especially to Bronwen Masemann for her invaluable editorial assistance.

REFERENCES

[1] Abu-Lughod, L. and Lutz, C.A. 1990. Introduction: emotion, discourse, and the politics of everyday life. *Language and the Politics of Emotion*. C.A. Lutz and L. Abu-Lughod, eds. Cambridge University Press. 1–23.

[2] Alashri, S., Kandala, S.S., Bajaj, V., Ravi, R., Smith, K.L. and Desouza, K.C. 2016. An analysis of sentiments on Facebook during the 2016 U.S. presidential election. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM) (Jul. 2016), 795–802.

[3] Andalibi, N. and Buss, J. 2020. The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks. *CHI '20: ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2020), 1–16.

[4] Andrejevic, M. 2007. Surveillance in the Digital Enclosure. *The Communication Review*. 10, 4 (Oct. 2007), 295–317.

[5] Andrejevic, M. 2011. The Work That Affective Economics Does. *Cultural Studies*. 25, 4-5 (Sep. 2011), 604–620.

[6] Arnold, M. 1960. *Emotion and Personality*. Columbia University Press.

[7] Bakir, V. and McStay, A. 2017. Fake News and The Economy of Emotions. *Digital Journalism*. 6, 2 (Jun. 2017), 154–175.

[8] Barocas, S. and Selbst, A.D. 2016. Big Data's Disparate Impact. *California Law Review*. 104, 3 (2016), 671–732.

[9] Barrett, L.F. 2017. *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt.

[10] Barrett, L.F. 2006. Solving the Emotion Paradox: Categorization and the Experience of Emotion. *Personality and Social Psychology Review*. 10, 1 (Feb. 2006), 20–46.

[11] Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M. and Pollak, S.D. 2019. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*. 20, 1 (Jul. 2019), 1–68.

[12] Ben Williamson 2019. Psychodata: disassembling the psychological, economic, and statistical infrastructure of "social- emotional learning." *Journal of Education Policy* (Oct. 2019), 1–26.

[13] Benjamin, R. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons.

[14] Birhane, A. and Guest, O. 2020. Towards decolonizing computational sciences. *arXiv*.

[15] Boehner, K., Boehner, K., DePaula, R., Dourish, P. and Sengers, P. 2005. Affect: From Information to Interaction. *CC '05: Proceedings of the 4th decennial conference on critical computing* (New York, NY, 2005), 59–68.

[16] Boehner, K., DePaula, R., Dourish, P. and Sengers, P. 2007. How Emotion is Made and Measured. *International Journal of Human-Computer Studies*. 65, (2007), 275–291.

[17] Boehner, K., Sengers, P. and Warner, S. 2008. Interfaces with the Ineffable: Meeting Aesthetic Experience on its Own Terms. *ACM Transactions on Computer-Human Interaction* 15,3 (December 2008). https://doi.org/10.1145/1453152.1453155

[18] boyd, D. and Crawford, K. 2012. Critical Questions for Big Data. *Information, Communication & Society*. 15, 5 (Jun. 2012), 662–679.

[19] Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A. and Van Bavel, J.J. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*. 114, 28 (Jul. 2017), 7313–7318.

[20] Brandt, M. and Stark, L. 2018. Exploring Digital Interventions in Mental Health: A Roadmap. *Interventions*. A. Shaw and D.T. Scott, eds. Peter Lang. 167–182.

[21] Brodny, G., Kolakowska, A., Landowska, A., Szwoch, M., Szwoch, W. and Wróbel, M.R. 2016. Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions. *HSI*. (2016), 397–404.

[22] Bucher, T. 2016. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20, 1 (Feb. 2016), 30–44.

[23] Burrell, J. 2016. How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society*. 3, 1 (Feb. 2016), 205395171562251–12.

[24] Cheney-Lippold, J. 2017. *We Are Data: Algorithms and The Making of Our Digital Selves*. New York University Press.

[25] Chui, S. 2020. A Comparative Study of the Interpretations of Emojis in Between U.S. and Chinese Users. *International Journal of Literature and Arts*. 8, 3 (2020), 108–11.

[26] Clynes, M. 1989. *Sentics: The Touch of the Emotions*. Prism Press Ltd.

[27] Crawford, K. et al. 2019. *AI Now 2019 Report*. AI Now Institute.

[28] Davies, W. 2017. How are we now? Real-time mood-monitoring as valuation. *Journal of Cultural Economy* 10, 1 (2017), 34–48.

[29] de Giustino, D. 2016. *Conquest of Mind: Phrenology and Victorian Social Thought*. Routledge.

[30] Desmet, P.M.A. and Roeser, S. 2015. Emotions in Design for Values. *Handbook of Ethics, Values, and Technological Design*. J. van den Hoven, P.E. Vermaas, and I. van de Poel, eds. Springer. 203–219.

[31] Dror, O.E. 2009. Afterword: A Reflection on Feelings and the History of Science. *Isis*. 100, 4 (2009), 848–851.

[32] Dror, O.E. 2001. Counting the Affects: Discoursing in Numbers. *Social Research*. 68, 2 (Jul. 2001), 357–378.

[33] Dror, O.E. 1999. The Affect of Experiment: The Turn to Emotions in Anglo-American Physiology, 1900-1940. *Isis*. 90, 2 (Jun. 1999), 205–237.

[34] Ekman, P. and Friesen, W.V. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*. 17, 2 (Feb. 1971), 124–129.

[35] Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

[36] Fernandez, L. and Matt, S.J. 2019. *Bored, Lonely, Angry, Stupid*. Harvard University Press.

[37] Fiesler, C. and Proferes, N. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*. 4, 1 (Mar. 2018), 205630511876336–14.

[38] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A.C. and Srikumar, M. 2020. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*. Technical Report #Berkman Klein Center Research Publication No. 2020-1. Berkman Klein Center for Internet and Society.

[39] Fridlund, A.J. 1994. *Human Facial Expression: An Evolutionary View*. Academic Press.

[40] Gates, K. 2011. *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*. New York University Press.

[41] Ghassemi, M., Celi, L.A. and Stone, D.J. 2015. State of the art review: the data revolution in critical care. *Critical care (London, England)*. 19, 1 (Mar. 2015), 118–9.

[42] Gould, D. 2010. On Affect and Protest. *Political Emotions*. J. Staiger, A. Cvetkovich, and A. Reynolds, eds. Routledge. 18–44.

[43] Greene, D., Hoffmann, A.L. and Stark, L. 2019. Better, Nicer, Clearer, Fairer. *HICSS 2019* (Jan. 2019), 2122–2131.

[44] Greene, G. 2020. *The Ethics of AI and Emotional Intelligence*. Partnership on AI.

[45] Hard Feelings — Inside Out, Silicon Valley, and Why Technologizing Emotion and Memory Is a Dangerous Idea: 2015. *https://lareviewofbooks.org/essay/hard-feelings-inside-out-silicon-valley-and-why-technologizing-emotion-and-memory-is-a-dangerous-idea*. Accessed: 2015-09-30.

[46] Hardt, M. 1999. Affective Labor. *boundary 2*. 26, 2 (1999), 89–100.

[47] Hassan, M.M., Alam, M.G.R., Uddin, M.Z., Huda, S., Almogren, A. and Fortino, G. 2019. Human emotion recognition using deep belief network architecture. *Information Fusion*. 51, (Nov. 2019), 10–18.

[48] Hearables Will Monitor Your Brain and Body to Augment Your Life: 2019. *https://spectrum.ieee.org/consumer-electronics/audiovideo/hearables-will-monitor-your-brain-and-body-to-augment-your-life*. Accessed: 2019-05-06.

[49] Heise, D.R. 2007. *Expressive Order*. Springer.

[50] Hochschild, A.R. 2003. *The Managed Heart: Commercialization of Human Feeling*. University of California Press.

[51] Hoey, J. Citizens, Madmen and Children: Equality, Uncertainty, Freedom and the Definition of State. SocArXiv 2021. https://osf.io/f463y/

[52] Höök, K., Isbister, K., Westerman, S., Gardner, P., Sutherland, E., Vasalou, A., Sundström, P., Kaye, J. and Laaksolahti, J. 2010. Evaluation of Affective Interactive Applications. *Emotion-Oriented Systems*. Springer Berlin Heidelberg. 687–703.

[53] Insel, T.R. 2017. Digital Phenotyping: Technology for a New Science of Behavior. *JAMA*. 318, 13 (Oct. 2017), 1215–1216.

[54] Isbister, K., Höök, K., Laaksolahti, J. and Sharp, M. 2007. The sensual evaluation instrument: Developing a trans-cultural self-report measure of affect. *International Journal of Human-Computer Studies*. 65, 4 (Apr. 2007), 315–328.

[55] Izard, C.E. 1972. *Patterns of Emotions: A New Analysis of Anxiety and Depression*. Academic Press.

[56] Jacobs, A.Z. and Wallach, H. 2019. Measurement and Fairness. *arXiv*.

[57] Jain, S.H., Powers, B.W., Hawkins, J.B. and Brownstein, J.S. 2015. The digital phenotype. *Nature Publishing Group*. 33, 5 (May 2015), 462–463.

[58] James, W. 1884. What is an Emotion? *Mind*. 9, (1884), 188–205.

[59] Jobin, A., Ienca, M. and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*. (Sep. 2019), 1–11.

[60] Johnson, D.G. 2018. AI, agency and responsibility: the VW fraud case and beyond. *AI & Society*. 0, 0 (Jan. 2018), 0–0.

[61] Karppi, T. 2018. *Disconnect: Facebook's Affective Bonds*. University of Minnesota.

[62] Kerr, I. and McGill, J. 2007. Emanations, Snoop Dogs and Reasonable Expectations of Privacy. *Criminal Law Quarterly*. 52, 3 (2007), 392–431.

[63] Kim, D., Frank, M.G. and Kim, S.T. 2014. Emotional display behavior in different forms of Computer Mediated Communication. *Computers in Human Behavior*. 30, (Jan. 2014), 222–229.

[64] König, A., Francis, L.E., Joshi, J., Robillard, J.M. and Hoey, J. 2017. Qualitative study of affective identities in dementia patients for the design of cognitive assistive technologies. *Journal of Rehabilitation and Assistive Technologies Engineering*. 4, (Jan. 2017), 205566831668503–15.

[65] Kramer, A.D.I., Guillory, J.E. and Hancock, J.T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*. 111, 24 (Jun. 2014), 8788–8790.

[66] Lazarus, R.S. 1991. *Emotion and Adaptation*. Oxford University Press.

[67] Leahu, L., Schwenk, S. and Sengers, P. 2008. Subjective Objectivity: Negotiating Emotional Meaning. *DIS '08*. (February 25-27 2008), 425–434.

[68] Leys, R. 2017. *The Ascent of Affect*. University of Chicago Press.

[69] Littlefield, M. and Johnson, J. 2012. *The Neuroscientific Turn: Transdisciplinarity in the Age of the Brain*. University of Michigan Press.

[70] Lively, K.J. and Heise, D.R. 2004. Sociological Realms of Emotional Experience. *American Journal of Sociology*. 109, 5 (Mar. 2004), 1109–1136.

[71] MacKenzie, D.A. 2016. *An Engine, Not a Camera*. The MIT Press.

[72] MacKinnon, N.J. 2020. Affect Control Theory Applied to Morality.

[73] Malin, B. 2014. *Feeling Mediated: A History of Media Technology and Emotion in America*. New York University Press.

[74] Martin, E. 2007. *Bipolar Expeditions*. Princeton University Press.

[75] McStay, A. 2018. *Emotional AI: The Rise of Empathic Media*. SAGE.

[76] McStay, A. 2016. Empathic media and advertising: Industry, policy, legal and citizen perspectives (the case for intimacy). *Big Data & Society*. 3, 2 (Sep. 2016), 205395171666686–11.

[77] McStay, A. and Pavliscak, P. 2019. *Emotional Artificial Intelligence: Guidelines for ethical use*. EmotionalAI.org.

[78] McStay, A. and Urquhart, L. 2019. "This time with feeling?" Assessing EU data governance implications of out of home appraisal based emotional AI. *First Monday*. (2019).

[79] Meyer, M.N., Heck, P.R., Holtzman, G.S., Anderson, S.M., Cai, W., Watts, D.J. and Chabris, C.F. 2019. Objecting to experiments that compare two unobjectionable policies or treatments. *Proceedings of the National Academy of Sciences*. 43, (May 2019), 201820701–6.

[80] Nagel, T. 1979. The Fragmentation of Value. *Mortal Questions*. Cambridge University Press. 128–141.

[81] Narayanan, A. and Vallor, S. 2014. Why software engineering courses should include ethics coverage. *Communications of the ACM*. 57, 3 (Mar. 2014), 23–25.

[82] Ngai, S. 2002. "A Foul Lump Started Making Promises in My Voice": Race, Affect, and the Animated Subject. *American Literature*. 74, 3 (2002), 571–602.

[83] Nissenbaum, H. 2010. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford Law Books.

[84] Olteanu, A., Castillo, C., Diaz, F. and Kiciman, E. 2016. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. (Dec. 2016), 1–44.

[85] Papacharissi, Z. 2014. *Affective Publics: Sentiment, Technology, and Politics*. Oxford University Press.

[86] Parisi, D. 2018. *Archaeologies of Touch: Interfacing with Haptics from Electricity to Computing*. University of Minnesota Press.

[87] Pasquale, F. 2015. Privacy, Autonomy, and Internet Platforms. *Privacy in the Modern Age The Search for Solutions*. M. Rotenberg, J. Horwitz, and J. Scott, eds. 165–173.

[88] Pasquale, F. 2015. *The Black Box Society*. Harvard University Press.

[89] Physiognomy's New Clothes: 2018. *https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a*. Accessed: 2018-02-03.

[90] Piantadosi, S., Byar, D.P. and Green, S.B. 1988. The Ecological Fallacy. *American Journal of Epidemiology*. 127, 5 (May 1988), 893–904.

[91] Picard, R.W. 2000. *Affective Computing*. The MIT Press.

[92] Picard, R.W. and Cosier, G. 1997. Affective intelligence — the missing link? *BT Technology Journal*. 15, 4 (Oct. 1997), 150–161.

[93] Picard, R.W. and Klein, J. 2002. Computers that recognise and respond to user emotion: theoretical and practical implications. *Interacting with Computers*. 14, (2002), 141–169.

[94] Picard, R.W. and Scheirer, J. 2001. The Galvactivator: A Glove that Senses and Communicates Skin Conductivity. *Proceedings from the 9th International Conference on Human-Computer Interaction* (New Orleans, LA, 2001), 1–6.

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

[96] Prinz, J.J. 2004. Introduction: Piecing Passions Apart. *Gut Reactions: A Perceptual Theory of Emotion*. Oxford University Press. 1–11.

[97] Reece, A.G. and Danforth, C.M. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Sci*. 6, 1 (Jul. 2017), 1–12.

[98] Reisenzein, R. 2009. Emotional Experience in the Computational Belief–Desire Theory of Emotion. *Emotion Review*. 1, 3 (Jun. 2009), 214–222.

[99] Ren, M., Nie, W., Liu, A. and Su, Y. 2019. Multi-modal Correlated Network for emotion recognition in speech. *Visual Informatics*. 3, 3 (Sep. 2019), 150–155.

[100] Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach: 2018. *https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election*. Accessed: 2018-03-18.

[101] Rhue, L. 2018. Racial Influence on Automated Perceptions of Emotions. (Dec. 2018), 1–11. http://dx.doi.org/10.2139/ssrn.3281765

[102] Russell, J.A. 2003. Core affect and the psychological construction of emotion. *Psychological Review*. 110, 1 (2003), 145–172.

[103] Russell, J.A. 2009. Emotion, core affect, and psychological construction. *Cognition & Emotion*. 23, 7 (Nov. 2009), 1259–1283.

[104] Scarantino, A. and de Sousa, R. 2018. Emotion. *The Stanford Encyclopedia of Philosophy*. E.N. Zalta, ed.

[105] Scheuerman, M.K., Paul, J.M. and Brubaker, J.R. 2019. How Computers See Gender. *Proceedings of the ACM on Human-Computer Interaction*. 3, CSCW (Nov. 2019), 1–33.

[106] Schlag, P. 1997. Law and Phrenology. *Harvard Law Review*. 110, (1997), 877–921.

[107] Schneider, S.M. and Morris, E.K. 1987. A history of the term radical behaviorism: From Watson to Skinner. *The Behavior Analyst*. 10, 1 (1987), 27–39.

[108] Schröder, T., Hoey, J. and Rogers, K.B. 2016. Modeling Dynamic Identities and Uncertainty in Social Interactions. *American Sociological Review*. 81, 4 (Jul. 2016), 828–855.

[109] Schuller, K. 2018. *The Biopolitics of Feeling*. Duke University Press.

[110] Shank, D.B. 2010. An Affect Control Theory of Technology. *Current Research in Social Psychology*. (2010), 1–13.

[111] Shank, D.B. 2014. Technology and Emotions. *Handbook of the Sociology of Emotions: Volume II*. Springer Netherlands. 511–528.

[112] Solomon, R.C. 2003. *What Is an Emotion?: Classic and Contemporary Readings*. Oxford University Press.

[113] Stark, L. 2019. Affect and Emotion in digitalSTS. *digitalSTS: A Field Guide for Science Technology Studies*. J. Vertesi and D. Ribes, eds. 117–135.

[114] Stark, L. 2018. Algorithmic Psychometrics and the Scalable Subject. *Social Studies of Science*. 48, 2 (Apr. 2018), 204–231.

[115] Stark, L. 2019. Facial recognition is the plutonium of AI. *XRDS: Crossroads, The ACM Magazine for Students*. 25, 3 (Apr. 2019), 50–55.

[116] Stark, L. 2018. Facial recognition, emotion and race in animated social media. *First Monday*. 23, 9 (Sep. 2018).

[117] Stark, L. 2016. *That Signal Feeling: Emotion and Interaction Design from Social Media to the "Anxious Seat."* Doctoral Dissertation, New York University.

[118] Stark, L. 2016. The emotional context of information privacy. *The Information Society*. 32, 1 (2016), 14–27.

[119] Stromfelt, H., Zhang, Y. and Schuller, B.W. 2017. Emotion-augmented machine learning: Overview of an emerging domain. *ACII*. (2017), 305–312.

[120] Terry, N.P. 2014. Big Data Proxies and Health Privacy Exceptionalism. 24 Health Matrix 65 (2014), 1–45.

[121] The Science of "Inside Out": 2015. *http://www.nytimes.com/2015/07/05/opinion/sunday/the-science-of-inside-out.html*. Accessed: 2015-07-06.

[122] Tomkins, S. and Izard, C.E. 1965. *Affect, Cognition, and Personality: Empirical Studies*. Springer.

[123] What happens when cars get emotional?: 2019. *https://www.fastcompany.com/90368804/emotion-sensing-cars-promise-to-make-our-roads-much-safer*. Accessed: 2021-01-17.

[124] Williamson, B. 2017. Moulding student emotions through computational psychology: affective learning technologies and algorithmic governance. *Educational Media International*. 10, 1 (Nov. 2017), 1–22.

[125] Wilson, E.A. 2010. *Affect and Artificial Intelligence*. University of Washington Press.

[126] Wilson, E.A. 2015. *Gut Feminism*. Duke University Press.

[127] Yadollahi, A., Shahraki, A.G. and Zaiane, O.R. 2017. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. *ACM Computing Surveys*. 50, 2 (Jun. 2017), 1–33.

[128] Zuboff, S. 2019. *The Age of Surveillance Capitalism*. PublicAffairs/Hachette.

[129] "Dehumanising, impenetrable, frustrating": the grim reality of job hunting in the age of AI: 2018. *https://www.theguardian.com/inequality/2018/mar/04/dehumanising-*

impenetrable-frustrating-the-grim-reality-of-job-hunting-in-the-age-of-ai. Accessed: 2018-03-05.
[130] "I created Steve Bannon"s psychological warfare tool': meet the data war whistleblower : 2018. *https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump*. Accessed: 2018-03-18.
[131] 2016. *Letter on Proposed Changes to the Common Rule*.
[132] 2018. The overly Social Network: Why the real villain in the Cambridge Analytica story might be Facebook | CBC Radio. (2018).
[133] 2007. *What is Emotion?* Jerome Kagan, Ed. Yale University Press.