Data Subjects' Conceptualizations of and Attitudes Toward Automatic Emotion Recognition-Enabled Wellbeing Interventions on Social Media

KAT ROEMMICH and NAZANIN ANDALIBI, University of Michigan, USA

Automatic emotion recognition (ER)-enabled wellbeing interventions use ER algorithms to infer the emotions of a data subject (i.e., a person about whom data is collected or processed to enable ER) based on data generated from their online interactions, such as social media activity, and intervene accordingly. The potential commercial applications of this technology are widely acknowledged, particularly in the context of social media. Yet, little is known about data subjects' conceptualizations of and attitudes toward automatic ER-enabled wellbeing interventions. To address this gap, we interviewed 13 US adult social media data subjects regarding social media-based automatic ER-enabled wellbeing interventions. We found that participants' attitudes toward automatic ER-enabled wellbeing interventions were predominantly negative. Negative attitudes were largely shaped by how participants compared their conceptualizations of Artificial Intelligence (AI) to the humans that traditionally deliver wellbeing support. Comparisons between AI and human wellbeing interventions were based upon human attributes participants doubted AI could hold: 1) helpfulness and authentic care; 2) personal and professional expertise; 3) morality; and 4) benevolence through shared humanity. In some cases, participants' attitudes toward automatic ER-enabled wellbeing interventions shifted when participants conceptualized automatic ER-enabled wellbeing interventions' impact on others, rather than themselves. Though with reluctance, a minority of participants held more positive attitudes toward their conceptualizations of automatic ER-enabled wellbeing interventions, citing their potential to benefit others: 1) by supporting academic research; 2) by increasing access to wellbeing support; and 3) through egregious harm prevention. However, most participants anticipated harms associated with their conceptualizations of automatic ER-enabled wellbeing interventions for others, such as re-traumatization, the spread of inaccurate health information, inappropriate surveillance, and interventions informed by inaccurate predictions. Lastly, while participants had qualms about automatic ER-enabled wellbeing interventions, we identified three development and delivery qualities of automatic ER-enabled wellbeing interventions upon which their attitudes toward them depended: 1) accuracy; 2) contextual sensitivity; and 3) positive outcome. Our study is not motivated to make normative statements about whether or how automatic ER-enabled wellbeing interventions should exist, but to center voices of the data subjects affected by this technology. We argue for the inclusion of data subjects in the development of requirements for ethical and trustworthy ER applications. To that end, we discuss ethical, social, and policy implications of our findings, suggesting that automatic ER-enabled wellbeing interventions imagined by participants are incompatible with aims to promote trustworthy, socially aware, and responsible AI technologies in the current practical and regulatory landscape in the US.

CCS Concepts: • **Human-centered computing** → Empirical studies in HCI.

Additional Key Words and Phrases: emotion recognition, affect recognition, artificial emotional intelligence, affective computing, emotion AI, wellbeing interventions, ethics, AI ethics, fairness, algorithmic accountability, social media

Authors' address: Kat Roemmich, roemmich@umich.edu; Nazanin Andalibi, andalibi@umich.edu, University of Michigan, Ann Arbor, MI, USA.

2573-0142/2021/10-ART308 $15.00
https://doi.org/10.1145/3476049

ACM Reference Format:
Kat Roemmich and Nazanin Andalibi. 2021. Data Subjects' Conceptualizations of and Attitudes Toward Automatic Emotion Recognition-Enabled Wellbeing Interventions on Social Media. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 308 (October 2021), 34 pages. https://doi.org/10.1145/3476049

1 INTRODUCTION

Human emotion holds powerful influence over how we engage with the world around us [68]. Emotions mediate our experiences and drive how we make decisions [52, 117]. Emotions are uniquely personal and private, yet vulnerable to manipulation [22].

Social media platforms provide distinctive contexts for people to share personal and emotional content, while also being personally and emotionally affected by interactions mediated on these platforms [156]. Social media companies have been implicitly or explicitly interested in emotions. For example, to better understand their users' emotions and whether and how they could shape them, in 2014, Facebook researchers conducted a large-scale experimental study to examine whether "emotional states could be transferred to others via emotional contagion" [115]. The public backlash to this study of emotion manipulation was widespread and severe [102, 132, 155]. Bottom-up criticism derived from news article commentators has demonstrated that the public had a variety of concerns about Facebook modifying their News Feed for emotional content and analyzing their subsequent engagement with the platform to infer its emotional impact, including concerns about being manipulated, being subject to research without consent, violation of expected use of data, and lack of trust in Facebook generally [94].

Despite negative public sentiment regarding technology companies engaging in the manipulation of and making inferences about an individual's emotion, as evidenced perhaps notably by Facebook's emotional contagion study, researchers and technology companies continue to deepen and expand the application of the growing emotion recognition (ER) market. ER, sometimes referred to as emotion AI or artificial emotional intelligence, is "achieved by the capacity to see, read, listen, feel, classify and learn about emotion life" enabled by "reading words and images, seeing and sensing facial expressions, gaze direction, gestures and voice…feeling our heart rate, body temperature, respiration, and the electrical properties of our skin, among other bodily behaviours" [130]. As costs for computing power continue to decline while advances in computational power rise [3], and sharing of personal and revealing information on social media continues to grow [144], development of ER applications on social media have spread across public and private sectors. Academic researchers are primarily interested in harnessing emotion data for public health purposes [51, 133], corporations use it to gauge opinion about their products and consumer preferences, and governments find it useful to understand public sentiment and assess security risks, to name a few.

Automatic ER-enabled wellbeing interventions rely on computational techniques to process the data a person generates in their day-to-day use of internet-connected devices to infer their emotions, and intervene accordingly. Data sources could include social media use, search engine use, wearable devices, voice assistants, and more. Computationally inferred emotions can be processed to make inferences and predictions about individual behaviors, medical and mental health conditions, and emotional states [89]. In the US, medical and mental health data is protected by federal and state legislation, but such digital inferences fall outside the scope of these protections [125]. As such, inferences of a person's emotions can be packaged and sold to insurance companies, advertisers, and other interested parties, often without that person's knowledge or consent [47, 89]. These inferences are of special interest to psychiatry and psychology — fields that have traditionally relied on self-reported patient surveys to diagnose conditions — as an intervention tool with potential to increase accuracy in patient diagnosis, detect conditions early, and identify people in moments of crisis or relapse [35, 103, 148].

There has been increasing enthusiasm for population scale research and monitoring, particularly in the medical and computational social science fields [16, 54, 122, 158, 163]. Social media platforms in particular are a uniquely rich source of emotional data, as individuals use these sites to disclose and disseminate sensitive information such as personal experiences and media content [19], as well as receive social support from friends within their network [25], with the potential to improve wellbeing [93]. In this study, we are interested in the application of ER for automatic wellbeing-related interventions on social media, which we define as the application of ER to automatically infer individuals' emotions and intervene accordingly, and to which we take a broad perspective, describing anything aimed at or framed as aiming to automatically provide support to social media users' wellbeing. We argue that individuals who post the emotional content on which automatic ER-enabled wellbeing interventions are trained and depend, and as potential subjects to those interventions, are stakeholders in the development and delivery of ER technologies. In our study, we center data subjects, which we refer to as individuals whose data enables ER and are possibly affected by its outcomes or applications.[1] Understanding the human impact of the surveillance, datafication, and commodification of data subjects' emotions is crucial to any evaluation of the ethical and responsible use of this emerging technology. Thus, in this study we seek to understand the attitudes of those data subjects who, by posting emotional content on social media, both contribute to the creation of and may be targeted by automatic ER-enabled wellbeing interventions.

Indeed, automatic ER-enabled wellbeing interventions on social media have been criticized in both academic scholarship and opinion pieces for the enormous harm they present to individual autonomy, individual privacy, and individual safety [33, 48, 53, 89, 125, 157]. However, the critical discourse surrounding this technology has grossly omitted engagement with data subjects, who are potentially affected by automatic ER-enabled wellbeing interventions and whose data make this technology possible, as a relevant social group [154] to gain their insights, perceptions, preferences, and attitudes toward automatic ER-enabled wellbeing interventions. Given the sensitive data that enables automatic ER-enabled wellbeing interventions [22], we submit that the participation of data subjects in its development, implementation, and delivery is critical to any potential ethical, trustworthy, and responsible implementation of this technology. In this study we contribute an in-depth understanding of data subjects' conceptualizations of and attitudes toward automatic ER-enabled wellbeing interventions on social media. We do not seek to make normative statements about the existence of automatic ER-enabled wellbeing interventions, but rather to center and promote the voices and concerns of data subjects. To that end, we conducted a series of in-depth semi-structured interviews with adult social media users in the United States who have experienced both positive and negative meaningful personal experiences in the past year and who reported having shared about them on social media.

Overall, we found that participants had negative conceptualizations of and attitudes toward automatic ER-enabled wellbeing-related interventions on social media, but in a minority of cases held more positive attitudes toward such wellbeing interventions when they targeted *other* individuals, rather than themselves. We first develop an understanding of why people held negative attitudes toward automatic ER-enabled wellbeing interventions. Negative attitudes were rooted in the way people compared traditional delivery methods of wellbeing interventions, by humans, to their conceptualizations of algorithmically-enabled wellbeing interventions, by AI. We identified four attributes that participants remarked were essential to supportive wellbeing interventions: 1) helpfulness and authentic care; 2); personal and professional expertise; 3) morality; and 4)

[1]The term *data subject* has been used by scholars to refer to individuals whose data enables technologies and who are impacted by it, although not always clearly defined [58, 105]. It also has a very different meaning in the General Data Protection Regulation (GDPR) context [14], which emphasizes identifiability, and is not how we use the term.

benevolence through shared humanity. Participants felt that these attributes could only be held by humans, expressing doubt that Artificial Intelligence (AI) could hold them. These comparisons between conceptualizations of automatic ER-enabled wellbeing interventions and human-delivered wellbeing interventions shaped attitudes toward automatic ER-enabled wellbeing interventions in a negative way.

We describe the tension between participants' negative conceptualizations of and attitudes toward automatic ER-enabled wellbeing interventions broadly, and how some participants imagined possible positive social benefits when realizing its potential, assumed impact on *others*. These participants conceived of automatic ER-enabled wellbeing interventions as a potential social good that could support academic research, increase access to wellbeing, and prevent egregious harm. Yet most participants maintained their negative attitudes toward automatic ER-enabled wellbeing interventions when conceptualizing its potential impact on others, revealing worries about the potential harm (e.g., re-traumatization, spread of inaccurate health information, inappropriate surveillance, interventions informed by inaccurate predictions) that automatic ER-enabled wellbeing interventions could cause others. Participants emphasized a requirement for individual and external control to potentially mitigate these harms. These observations highlight the importance of including data subjects in emerging technologies' development rather than conceiving of data subjects as "others" and *assuming* what *their* wellbeing entails. Finally, we discuss qualities in either the development or delivery of the data subjects' conceptualized intervention upon which data subjects' attitudes were dependent. These qualities include: 1) accuracy; 2) contextual sensitivity; and 3) positive outcome.

We then situate these findings within the discourse surrounding policy and ethical implications of automatic ER-enabled wellbeing interventions. We argue that in the current practical and regulatory landscape in the US, automatic ER-enabled wellbeing interventions are incompatible with ethical and socially responsible AI applications. Further, we express concern for current state social media intervention processes that include police intervention for mental health crises, especially for racial/ethnic minority populations in the US. We speculate (and critique) that instead of honoring the concerns of data subjects, the entities that employ emotion surveillance technologies have focused on promoting a rhetoric of "safety through surveillance," and shaping social norms to accept ubiquitous surveillance under the guise of public safety.

The constant monitoring of data subjects' emotions carries tremendous risk in its power to shape human behavior. By developing automated and predictive systems built on normative assumptions of human emotion, new realities are built around the expectations and anticipations of their outcome while perpetuating stigma around emotion and mental illness [142]. As warned by Couldry and Mejias, "the constant watchability of our every thought and action by external forces changes the field of power in which we exist, transforming a supposed order of individuals into a collection of living entities plugged into an external system" [58]. In other words, even when and if individual people are not subject to the gaze of surveillance capitalism, its practice holds implications for all individuals as part of a larger system [58]. The "normative weight" of the models used to develop ER technologies, what data is collected to feed those models, and what people (and society) believe about the inferences and predictions of the "interiority, judgments, and potential future actions of human beings" matters if we are to understand the ethical and social implications of ER [159]. We align our position with these works, and center the data subject to contribute to understanding ER's impact on humans and society at large.

2 PRIOR WORK

In this section, we summarize the historical development of ER in computing, discuss the current state of automatic ER-enabled wellbeing interventions, and document the ethical concerns

surrounding this technology, showing how our work is partly motivated by the absence of data subjects' voices in these valuable debates.

2.1 Automatic Emotion Recognition

We first summarize the history of ER technologies that enable automatic ER-enabled wellbeing interventions. Automatic ER-enabled wellbeing interventions use emotion AI to infer, detect, predict, and recognize emotions through the surveillance of an individual's everyday use of internet-enabled devices, classify those emotional inferences, and respond in a personalized way [130, 168]. Data used to computationally infer emotion include emotional content explicitly or implicitly disclosed on social media [23, 27, 114], search engine logs [91], sensor data [147, 150], and voice data [123], to name a few. ER is foundational to computational wellbeing interventions, as "an affect-sensitive interface can never respond to users' affective states if it cannot sense their affective states" [43]. The theoretical frameworks underpinning ER are varied and multidisciplinary, rooted primarily in the works of 19th century American philosopher and psychologist William James, who proposed that emotions were secondary to and embodied by perceptions of physiological changes [109], and British naturalist and biologist Charles Darwin [60], whose study of facial and bodily expressions, understanding of emotions as universal, and treatment of emotions as discrete entities deeply influenced modern American psychologist Paul Ekman's understanding of emotion [71]. Ekman developed a theory of six basic emotions — anger, disgust, fear, joy, sadness, and surprise — informing much of how the field of Affective Computing conceptualizes emotions [70, 72, 145]. Some have criticized Affective Computing for its focused attention to technical challenges rather than broadening its theoretical underpinnings, which have in some cases shown to be in conflict with developments in affective science [39, 43, 153]. For example, some affective neuroscience research has shown that subtle emotional behavior may not be explicitly processed by the person exhibiting the behavior, challenging traditional assumptions of emotion [42]. Others have suggested Ekman's theory of basic emotions is too definitionally rigid [151] and might require reformulation in light of neuroimaging data that complicates this view [45].

This section highlights the physiological assumptions of emotion that have shaped computational methods to detect, infer, and predict emotion.

2.2 Automatic Emotion Recognition-enabled Wellbeing Interventions on Social Media

Social media in particular has shown promise as platforms from which to harvest emotion-intensive data [15, 48, 62, 65, 114, 124, 148], and infer mental health conditions such as schizophrenia [37, 134], depression [54, 64, 148, 164], post-partum depression [63], or post-traumatic stress disorder [55]. Researchers across disciplines, from medicine to computing, consider the potential benefit of computationally inferring emotions to promote public health [75] by way of early diagnosis of illness [149], sentiment detection and behavior surveillance [97, 136], and real-time intervention [51, 90, 135, 161].

Due to the intimate way in which many people use social media, social media is considered both as a suitable source from which to infer emotions, and as an ideal platform to *intervene* based on those inferences of emotion. Opinion pieces regarding automatic ER-enabled interventions have offered mixed support and criticism, with some questioning its ethical and privacy implications while others laud the interventions' support of suicide prevention efforts [31, 131, 138, 140]. Perhaps the most prevalent example of an automatic ER-enabled wellbeing intervention is Facebook's suicide prevention intervention, which uses a combination of n-gram based linear regression and DeepText-based neural network models to flag users at risk of imminent harm, and intervenes by suggesting the contact number of the National Suicide Prevention Lifeline and offering the ability

to chat with a crisis worker; the case is also sent for review by a human reviewer, who then decides if the company will involve police for a welfare check [34, 69, 90].

While prior work has examined peoples' attitudes toward the development, use, and implications of ER or similar approaches on social media generally [22, 49, 76, 78], relatively little work has explored peoples' attitudes and conceptualizations of the application of ER inferences to develop, implement, and deliver automatic ER-enabled wellbeing interventions (a key application domain for ER technologies). Beyond general attitudes toward ER, it is important to examine peoples' attitudes toward ER's various applications for particular purposes and in specific contexts such as for purposes of wellbeing (our focus) or advertising, and in contexts such as social media (our focus), the workplace, or education. The various ways in which automatic ER-enabled wellbeing interventions have been researched and deployed, as reviewed here, point us to a lack of including data subjects in the development and deployment of such interventions.

2.3 Automatic Emotion Recognition-enabled Wellbeing Interventions: Ethics and Values

The growing interest and development of algorithmically inferred emotions and associated interventions has raised new ethical questions and considerations in the areas of privacy, harm to vulnerable populations, transparency, and fairness. Echoing past work that has shown predominantly negative data subjects' attitudes toward ER broadly, finding that people perceive algorithmic inferences of emotion as invasive and intrusive [22, 85], scholars from many disciplines including law, computing, philosophy, and psychiatry have sounded the alarm on the potential of targeted wellbeing interventions to infringe on individual privacy and autonomy [33, 48, 53, 89, 125, 157]. Additionally, scholars have warned of ER's potential to increase harm to vulnerable mental health patients through amplification of mental health bias and potential misuse of data [48, 59, 119, 125], express concern about ER's lack of algorithmic transparency [31, 34, 48, 59, 87, 92, 113], and raise doubt in ER's algorithmic fairness [29, 31, 48, 87, 92] and testing practices [47, 48].

While ethical and privacy implications of automatic ER-enabled wellbeing interventions have been discussed across disciplines, there has been little scholarly engagement with the data subjects whose data enable, and who are potentially affected by, these interventions. In their study of social media users' attitudes toward ER algorithms broadly, Andalibi and Buss discovered that people often had "negative reactions to ER using social media data"[22]. Participants in their study expressed feelings of distrust toward social media companies, and were concerned with how the social media company would use their emotional data [22]. They felt that social media companies did not regulate their algorithms, and did not trust social media companies to take responsibility for the algorithm's consequences [22]. Ford et al. found similar results in their survey of user perceptions toward Facebook using emotion data to provide targeted mental health advertising. Participants in their study were not comfortable with their Facebook posts being analyzed for targeted advertising by algorithms, and even more uncomfortable with their posts being analyzed by human reviewers [85]. Studying the context of digital phenotyping by technology companies broadly, Costello and Floegel found that individuals with mental illness were wary of automated assessments of mental health and mood-tracking applications. The participants in their study were concerned about the profit motives behind such applications, and were distrustful that their personal data would be used responsibly [57]. These studies highlight an overall public distrust of social media companies collecting, processing, and sharing sensitive information such as emotion data [22, 85].

To date, empirical research has only sought either a very broad [22] (i.e., ER in general) or a limited [85] (i.e., targeted mental health-related ads) understanding of data subjects' attitudes toward ER use on social media. A notable application of ER is that of wellbeing-related interventions. Thus, we

build on these studies, contributing an in-depth understanding of data subjects' attitudes toward an unrestrained imagination about the development, implementation, and delivery of automatic ER-enabled wellbeing interventions on social media. We center the preferences, needs, and attitudes of data subjects in a discussion of the ethical, social, and policy implications of automatic ER-enabled wellbeing interventions on social media.

3 METHODS

3.1 Recruitment

We conducted in-depth, semi-structured interviews (N=13) lasting between 77 to 120 minutes (average=106 minutes) with adult social media users in the US. We recruited participants via a screening survey and conducted interviews over voice and/or video call. We transcribed the interviews for analysis. We shared calls for participation via personal social media, personal networks, and Craigslist. We chose Detroit and Houston Craigslist pages in an effort to achieve a diverse participation pool, in consideration of these cities' high racial/ethnic minority populations [8, 9]. In three cases, the interview participant was acquainted with the interviewer. To preserve the integrity of the data, another researcher on the team conducted the interviews in those three cases. Participants received a $30 honorarium. This study was approved by our institution's IRB.

3.2 Participation

Out of 100 responses to the screening survey, we invited 20 to participate in the interviews. Survey respondents who did not meet the minimum criteria (based on age, location, and behavior) did not proceed to the next step of the survey. Out of 20 invited to interview, 13 signed a consent form, scheduled, and appeared for the interview. Survey questions included inquiries regarding social media usage, such as whether they had shared positive and negative personal experiences on social media in the past year. Decisions to invite respondents to the interviews were conducted in an iterative manner and partly made based on the identities and experiences represented by the data collected by that point in time. We aimed to interview people who had both positive and negative emotional experiences, and shared about them in some form on social media, due to our study's goal of capturing conceptualizations of and attitudes toward emotion inferences based on real experiences and posting behavior, and our focus on emotions. These real experiences provided a basis for our participants to draw from when probed for scenarios designed to elicit their values and imaginaries regarding automatic ER-enabled wellbeing interventions on social media.

In addition, our goal also included capturing a range of identities (e.g., race/ethnicity, age, gender) and experiences. Examples of positive experiences represented included career accomplishments, educational attainment, and home ownership. Examples of negative experiences represented included job loss, health concerns, and relationship complications. Our study's racial/ethnic makeup included one Indian, two Asian, two Black, and eight white participants. Ages of participants ranged from 22 to 58, with an average age of 32.4. Gender identifications included nine women, one man, one gender-fluid, one agender, and one genderqueer. Education completed included five participants with college degrees, six with graduate degrees, one with some high school, and one with some college. Eleven out of thirteen participants used Facebook regularly. Other social media used included Facebook groups, Instagram, Linkedin, Twitter, Tumblr, AO3, Reddit, Snapchat, Twitch, YouTube, and Discord.

3.3 Interviews and Scenarios

3.3.1 Interviews. We followed a semi-structured protocol when conducting interviews to allow for exploration and flexibility. Interviews started by asking participants about their social media use,

social media sharing behaviors (particularly in regard to meaningful and emotional experiences), understanding of what happens to such data when shared, and expectations for privacy in those contexts. To facilitate recall, we probed interview participants with what they had shared with us in the survey when needed (e.g., *"you had mentioned..."*) and encouraged them to refer to their posts as we spoke if they wanted. By eliciting recall of specific experiences in the first phase of the interview, we were able to better understand how participants used social media to share emotional and personal experiences, and positioned participants in a context of emotion-situated experiences when exploring scenarios in the next phase. We did not observe any struggle with participants recalling experiences.

3.3.2 Scenarios. The next phase involved using speculative scenarios to elicit values, concerns, and attitudes toward ER on social media. Participants were allowed flexibility as to which experience they wanted to discuss during the scenarios.

Scenarios have been used in prior HCI and CSCW work. Though a complete review is outside our scope, we emphasize our methodological choice to use speculative scenarios due to its helpful application eliciting values toward technologies (and especially emerging technologies) [18, 40, 44, 99, 167] in cases where people may not be familiar with the technology or topic being examined [83]. Other HCI research has used scenarios to develop theory rather than assess values toward technology [24].

To probe for peoples' values and conceptualizations of automatic ER-enabled wellbeing interventions, we first probed for values related to entities making inferences or predictions based on emotional content shared on social media, and then probed to ask how they might feel if those inferences or predictions were used to offer *"wellbeing support or help them feel better."* We kept this question broad, because our goal was to understand what participants would imagine these interventions to be like, or what examples from their experiences they might share with us. Despite critique of using scenarios based on the presumption that what people will do is different than what they say they will do when imagining the scenario, past work has shown that in emotional contexts, such as those used in this study, people tend to respond similarly to scenarios as they would in real life [104]. Further, our goal was to surface participants' concerns and attitudes, for which hypothetical scenarios are useful tools [167].

Our use of scenarios is informed by prior work in algorithmic folk theory and privacy, which suggests that understanding what people *think* technology (and algorithms) can do or already do, is just as important as understanding how the technology operates in practice [74, 166]. Building on these works, [21, 74, 166], our study centers data subjects' imaginaries of algorithms and uses hypothetical scenarios to probe for human values including and beyond data privacy. Our focus in this work is data subjects' conceptualizations of and attitudes toward automatic ER-enabled wellbeing interventions, and the factors we found to shape those attitudes — not all the collected data.

In using scenarios, participants were asked to consider the positive and negative experiences they had posted about in the past year (an inclusion criterion for the study), then imagining how they would feel *"if the social media site on which [they] shared their experience had used computational methods to infer [their] emotions, either at the time or after [their] posting."* We then asked participants questions regarding their attitudes and values toward these emotion inferences, based on their social media data. Additionally, we asked participants to consider two specific applications of inferences made from their emotion data: advertisements (not our focus here) and wellbeing-related interventions. Questions asked in these contexts were used to determine factors that shaped participant attitudes.

3.3.3 Scenario Specifics. Scenarios were presented to participants via a link to a Google document. We randomized the order in which scenarios were presented to participants. The document included the following text, once for positive and once for negative emotional experiences, as determined by the participants themselves:

'I would like you to think about something [positive/negative and personal] that brought out [positive/negative] emotions for you. Maybe the experiences we talked about earlier. Now consider this scenario: You had shared on [insert social media they use most] about that, and had explicitly shared how you felt about it. Everyone reading it would have been able to understand what your experience was and how you felt, there was no ambiguity. Now imagine that [insert social media they posted on] used computational methods to detect what emotions you felt at the time of posting that.'

We began with one experience, and if appropriate and time permitting, we asked *"How about if this was related to your other experience?"* Participants were asked to share which experiences they thought about in relation to the scenarios so as to provide context and establish what emotional connections they made. For all cases, these included the emotional experiences participants shared in the screening survey and in the initial phase of the interview; sometimes participants brought up new topics. Once the emotional context of participants' imagining of the scenario was established, we probed to elicit their attitudes, concerns, and reactions toward algorithmic inferences of emotion based on social media data. This paper's focus is not these general emotional inferences, but toward automatic ER-enabled wellbeing interventions specifically, so we do not provide additional detail.

We then proceeded to ask participants questions about prediction, such as *"How do you feel about your post being used to predict how you might feel in the future? Tell me more about that. Why do you think companies might do that? How do you feel about that?"* Specific to this study, we then asked questions like: *"How do you feel about the platform using this prediction or detection to intervene in some way to support your wellbeing or help you feel better?"* We probed for other application domains of ER using social media data, but detail is not provided here as their scope is beyond the focus of this paper. We were intentionally broad when conducting scenarios regarding automatic ER-enabled wellbeing interventions. As emphasized throughout this paper, we take a broad perspective to automatic ER-enabled wellbeing interventions on social media, describing anything aimed at or framed as aiming to automatically provide support to social media users' wellbeing.

Our focus was not so much the particularities of the emotional experience, but more so how participants felt about that data being fed into ER algorithms to be used for wellbeing interventions, and attitudes toward those resulting interventions themselves. Sometimes, if we needed to probe more, we brought up another example that they had mentioned earlier in the interview or survey, and asked the same questions to reveal participants' attitudes toward automatic ER-enabled wellbeing interventions.

3.4 Analysis

Interviews were transcribed to enable qualitative analysis. We analyzed the data using open coding followed by axial coding [56]. The second author and the interviewer researcher engaged in weekly meetings to discuss observations and identify potential codes and patterns surfacing during the interviews and took detailed notes. These frequent discussions informed our data collection efforts, meaning we attempted to recruit individuals who may have different and similar perspectives based on the screening survey responses. We followed by formal open coding. The interviewer researcher team member first open coded five interviews. The second author and that team member then discussed each code and associated data in detail, refined codes, and grouped them into larger initial themes. The same team member then coded another five interviews and grouped codes into new themes or ones already developed, followed by coding the remaining interviews (we identified

no new themes in this last phase). The team member and the second author engaged in weekly discussions to reflect on and refine identified themes and draw connections between them, as well as to identify further points of probing for future interviews. In this paper, we report on themes related to our focus, conceptualizations of and attitudes toward automatic ER-enabled wellbeing interventions on social media, not all the themes we discovered in the data. Once we were done with analyzing the 13 interviews, we stopped recruiting more participants because we had begun to hear similar narratives, as identified through weekly discussions between researchers and as confirmed by our analysis. It is important to note that scenarios were not intended to be used in an experimental setting, but more so used as probes, and so it was not conducive to our analysis to associate participants' responses to specific details of scenarios.

3.5 Limitations and Opportunities

Our study's goal was neither representation nor generalizability [7]. Indeed, the demographics in our study are unique in some regards. We understand that a study about sharing emotional experiences on social media may not have elicited high participation among male-identifying individuals [61], and our sample thus included majority women and other underrepresented genders. Additionally, our study is in alignment with other studies of emerging technology [17, 95] in that most of our participants have attained at least a college degree and therefore may have been more familiar with technology than the general population. Despite these unique demographics and limitations, our work provides unique insights into conversations regarding emerging technologies. Future work on attitudes of automatic ER-enabled wellbeing interventions should include people with lower educational attainment, older adults, children, diverse races/ethnicities, those with mental illnesses, and people in diverse cultures and geographic locations. Future work can also use methods such as large scale surveys to examine generalizability of our findings.

In-depth interviews with smaller sample sizes allow researchers to make interpretive and generative conclusions rather than conclusions that are definitive and generalizable. Diligent participation selection allows us to explore topics of interest in depth. Our confidence in the validity of reported themes is high, as narratives were similar throughout data collection, confirmed by our analysis.

Furthermore, some participants expressed using privacy settings in general; therefore, it is possible that our work suffered from self-selection bias. Nonetheless, despite these concerns our participants had *still* chosen to share about emotional experiences on social media, which was an inclusion criterion for our study, as it enabled participants to engage in the scenarios around which our study was designed. Our study's goal was primarily to understand and make sense of how people that share emotional content on social media construct meaning from automatic ER-enabled wellbeing interventions, and what values and concerns they hold in this context. Though some participants may have had imperfections in recalling their past experiences sharing on social media, this imperfection would not have interfered with this goal.

Of course, those who do not post emotional content can also be subject to ER and interventions and engaging with them is important for future work. Yet, as a first step, we wanted to have our understanding grounded in participants' conceptualizations of emotions and emotional experiences, thus our choice of sampling.

For future work, we especially emphasize the importance of examining attitudes toward automatic ER-enabled wellbeing interventions for specific mental health conditions. Individuals living with mental illnesses are not a monolith, and attitudes toward automatic ER-enabled wellbeing interventions may differ across and within subgroups of people with mental illness. Prior research has analyzed, for example, how people with eating disorders share supportive and intimate content on social media, and how they might be impacted by the *coded gaze* that makes possible algorithmic inferences of user behavioral state and inferences linked to content moderation [50, 77, 79, 80, 143].

Future work could consider how and if data subjects' use of social media for social support in mental health or other emotion-situated contexts shapes their attitudes toward automatic ER-enabled wellbeing interventions that target their condition specifically.

4 RESULTS

We discuss data subjects' conceptualizations of and attitudes toward automatic ER-enabled wellbeing interventions on social media. As a reminder, by interventions we take a broad perspective, describing anything, automatic and ER-enabled, aimed at or framed as aiming to provide support to social media users' wellbeing, which necessarily includes the inference of emotions as a basis. We probed for participants' imaginaries of 1) ER-enabled inferences feeding into automatic ER-enabled wellbeing interventions, and 2) automatic ER-enabled wellbeing interventions. We avoided describing or prescribing an understanding of these phenomena for participants. Our findings suggest that social media users have predominantly negative attitudes toward automatic ER-enabled wellbeing interventions on social media. First, we discuss data subjects' negative conceptualizations of automatic ER-enabled wellbeing interventions. Next, we discuss how people imagined the impact of automatic ER-enabled wellbeing interventions on others. A minority of participants felt tension between their broad negative attitudes toward automatic ER-enabled wellbeing interventions and their conceptualizations of them as a potential social good for others. Most participants however maintained their general negative conceptualizations of ER-enabled wellbeing interventions when thinking of its impact on others. Participants expressed concern for the harm automatic ER-enabled wellbeing interventions may pose, and stressed that individual people should have control over whether they would be subject to them. Lastly, we discuss qualities upon which participants' attitudes depended.

4.1 Broad Conceptualizations of Emotion Recognition-enabled Wellbeing Interventions: Human versus AI

The majority of participants held negative attitudes toward automatic ER-enabled wellbeing interventions. We found that these attitudes stemmed from participants' comparisons between current state wellbeing interventions, delivered by humans, and what they imagined to be future state wellbeing interventions, enabled and delivered by ER technologies. The human versus AI dichotomy was a prevalent theme in participants' conceptualizations, as they considered whether AI could hold certain attributes they considered to be held by humans in wellbeing-supportive roles. These attributes included: 1) helpfulness and authentic care; 2); personal and professional expertise; 3) morality; and 4) benevolence through shared humanity.

4.1.1 Helpfulness and Authentic Care. Some participants doubted the helpfulness of automatic ER-enabled wellbeing interventions, and their ability to deliver authentic care. P1, who had personal experience with mental illness, and drew upon that when discussing their attitudes toward automatic ER-enabled wellbeing interventions, reflected on past experiences searching for suicide-related information on Google, saying: *"If you Google like how to kill yourself or whatever, or Google automatically served you just like the 1-800 like suicide hotline number, that as someone who had been suicidal did not strike me as very effective."* P1 later remarked: *"I don't know that a computer is able to serve the right information to help someone*," illustrating skepticism about algorithms' ability to provide information that would be helpful to individuals in need of support in moments of distress and vulnerability. Other participants signaled their need for wellbeing support to feel authentic, and felt uncertain that an automatic ER-enabled wellbeing intervention could provide authentic and thus helpful support. On lack of perceived genuine care, P5 said: *"People are people and an algorithm is an algorithm, right? It's not looking to read and ignore like most people. I make a private*

post on Tumblr, pretty much everybody either just casually hearts it to let you know they're there or ignores it completely because that's uncomfortable. But the algorithm is not there out of any form of interpersonal care, even if it's been put there by a human being. I don't know if I could ever envision a world in which it was put there to genuinely help people, which is me being a real cynic but why would they care? I don't know." These examples show how automatic ER-enabled wellbeing interventions can feel impersonal and unhelpful compared to interventions authentically delivered by caring, trained human professionals. Further, the unique insights provided by P1, who disclosed having a mental illness, point to a need to better understand the attitudes of those that live with mental illness toward automatic ER-enabled wellbeing interventions in future work.

4.1.2 Personal and Professional Expertise. Participants' remarks reflected a belief that automatic ER-enabled wellbeing interventions lack the expertise that humans have, either due to 1) their professional training or 2) personal experiences. On the first, participants compared automatic ER-enabled supportive interventions to humans trained to provide expert support and interventions to people in their community in times of distress (e.g. mental health professionals), arguing that AI does not, and cannot, compare to expertly trained and trusted professionals due to their expertise (rather than the ability to care as described in 4.1.1). For example, P3 said: *"I don't know, a therapist went to grad school for it. They've studied the thing."* Echoing this sentiment, P10 said: *"I don't think that's appropriate...because I think it takes a lot of information and often a medical professional to let someone know if they're going through a particular, like a clinical problem, or if they're likely to have a clinical problem in the future."* Believing that algorithms cannot be as expert as humans would be in providing supportive interventions was a significant factor contributing to negative attitudes toward automatic ER-enabled supportive interventions on social media.

On the latter, participants spoke of the enormous trust people put in friends and other community members to help in times of crisis, and could not imagine themselves able to trust in automatic ER-enabled wellbeing interventions to meet that need. In comparing trained mental health professionals to AI, P3 remarked: *"They also have a certain amount of community attached. I don't feel like an AI could get there."* P3 continued: *"No, there's a reason why you might sad Tweet about things but you, in the end, will still rather call a friend and talk about it."* Participants had a difficult time imagining a space where automatic ER-enabled wellbeing interventions would be welcomed, as their need for support was already filled by empathetic and compassionate humans with whom they were personally connected. Ultimately, most participants struggled to imagine how algorithmic interventions could help in the same way humans do — supporting each other within one's personal network and community — and whether there was a need for automatic ER-enabled wellbeing interventions at all. We note that the predominantly college-educated demographics in our sample may have influenced this view, as college-educated persons tend to have more robust networks of support [116]. In section 4.2, "Negative Conceptualizations of and Attitudes Toward Automatic ER-enabled Wellbeing Interventions for Others," we discuss conceptualizations some participants had that automatic ER-enabled wellbeing interventions could be a benefit to people without support from family and friends.

4.1.3 Morality. Participants were skeptical of automatic ER-enabled wellbeing interventions on social media due to their underlying assumptions about and associated attitudes toward social media platforms' financial motivations. Those that traditionally deliver wellbeing interventions such as mental health practitioners are held to certain ethical standards of conduct, establishing an expectation of ethical and moral practice that engenders trust in those receiving wellbeing support or intervention. However, when participants imagined automatic ER-enabled wellbeing interventions on social media, they were cynical that algorithmic interventions — or the platforms that deliver them — could have or prioritize morality or ethics due to the assumed financial incentives of the

companies that own the intervention. For example, P5 voiced their cynicism in the intervention's moral and altruistic intentions: *"[I]t could be 100 percent innocent, people who want to make people feel better, but I'm also a bit of a pragmatic, realistic person and I know that there's money in it, and they'll do it for money regardless of where the idea originated or where it came from."* P2 elaborates further, explicitly questioning the ethical incentives of such interventions: *"I'm just not so convinced that the financial incentives of the companies are such that ... ethical incentives would take priority."* P9 echoed this sentiment: *"I don't think they could provide support to us to feel better. I think like they just want us to, I think their goal is to earn money."* P9 reflected that while some kind of support may be nice when one is feeling lonely, the financial motivations of automatic ER-enabled wellbeing interventions could taint how they view, and therefore receive, that intervention: *"I think it could make us feel nice I guess. But if there is a way that they are going to sell us a product, I think that would change how we view, how we see them...At the end of the day it's not our family, our friends, so it's not like genuine care. It's just trying to sell you something."* These sentiments illustrate the important role of social media users' cynicism toward companies' moral intentions in shaping attitudes toward supportive interactions companies may provide. Moreover, these accounts highlight participants' discomfort of their emotions and vulnerability being commodified and capitalized by platforms delivering automatic ER-enabled wellbeing interventions.

4.1.4 Benevolence through Shared Humanity. Participants doubted whether they would welcome automatic ER-enabled wellbeing interventions in the same way as they would from another human. In contrast to the benevolent disclosures and interventions that take place in the context of a trusted relationship with a mental health professional or close friend, participants felt that interventions delivered through algorithmic means were intrusive and creepy, with suspect intentions. As P13 said: *"If it was something about like being sick or something, I don't know. In one case, I feel like maybe it would be good because maybe that will push you to go get it checked out, but at the same time, I'm like, that's kind of...I don't know. Maybe going a little too far. Maybe it's a little too intrusive."* Similarly, P5 said: *"I do certainly think there is a positive way in which that system could be used. I still think it's kind of creepy but...there isn't an innocence in that sort of concept or an idea."* Thus, although some participants charitably acknowledged positive possible uses for automatic ER-enabled wellbeing interventions, they resolved that those possibilities did not outweigh their perceptions of the interventions' intrusiveness and creepiness or skepticism of its intentions.

Participants felt that humanity was an essential and primary attribute required of supportive wellbeing interventions, doubting whether non-human, automatic ER-enabled wellbeing interventions could hold secondary attributes that they felt traditionally human-delivered wellbeing interventions held, such as morality, helpfulness, and authentic care. For example, P3 feared that because the algorithms that would deliver interventions (despite the humans that wrote them) do not share humanity with humans, they therefore lack attributes such as morality that otherwise make humans more resistant to manipulation by bad actors: *"But an algorithm is a thing. It's not a person and it doesn't have wants or desires or anything. It isn't similar to you in the way that you both have a shared humanity. It's more of a thing, and that makes me uneasy. Because that means that thing in the wrong hands can do a lot of damage. It's not a person."* Participants felt supportive wellbeing interventions required an element of humanity that non-human algorithms and social media platforms cannot provide. P1 remarked: *"It feels really impersonal. I don't know. I think it takes more, I think it takes real empathy from a real person as opposed to some generic advice and I don't think just giving someone a 1-800 number or even just talking to a stranger on suicide hotline is really the best intervention long term."* Echoing this sentiment, P3 said: *"...it's good to be accurate [with recognizing and predicting emotions], but there's no humanity in it, right?"* In these examples, participants felt that because automatic ER-enabled wellbeing interventions are computationally

derived, they lacked the essential attributes of humanity and personhood that would help the recipient receive the intervention in a way they only could if delivered by a human. In the end, participants were resistant to an algorithm's ability to embody humanity and were doubtful that automatic ER-enabled wellbeing interventions would be helpful to whom they (are framed to) intend to support.

4.2 Positive Conceptualizations of and Attitudes Toward the Impact of Automatic Emotion Recognition-enabled Wellbeing Interventions for Others

Some participants imagined latencies of benefit and harm differently when conceptualizing the impact of automatic ER-enabled wellbeing interventions on others compared to their predominantly negative attitudes about the technology in general. A minority of participants, with reservation, imagined automatic ER-enabled wellbeing interventions as a potential social good. Relative to their generally negative attitudes toward automatic ER-enabled wellbeing interventions, these participants responded more positively to the idea of automatic ER-enabled wellbeing interventions on social media when conceptualizing it as a tool that could benefit *others,* rather than themselves. For these participants, automatic ER-enabled wellbeing interventions held potential as a social good that could benefit others in the following ways: 1) by supporting academic research; 2) by increasing access to wellbeing support; and 3) through egregious harm prevention. It is important to note that we did not ask about these potential positives (or negatives) directly; rather these came up organically as participants discussed their attitudes and were probed to share details. In these examples, the participants maintained their negative attitudes generally, but held slightly more positive attitudes toward the specific use case of promoting social good more broadly. Even in cases of automatic ER-enabled wellbeing interventions for social good, participants expressed caveats that collection of emotion data and subsequent potential interventions should be transparent to individuals, and that they should meaningfully consent to it.

4.2.1 Supporting Academic Research. Several participants noted that automatic ER-enabled wellbeing related interventions could be used to support researchers. For example, P7 said *"I would want that to be used in research, and in mental health studies."* Some felt it was best that the intervention would be developed and implemented by researchers. For example, P3 suggested they *"could trust a brand new person creating this new app with neuroscience and psychiatrist research that has the data to be like, 'Oh yeah, I think this is going to help change the world.'"* In this example, the participant was generally cautious about social media's deployment of automatic ER-enabled wellbeing interventions, but felt enthusiastic about the potential of more trusted academic researchers to create them. Other participants who were resistant overall to emotion data collection supported the idea of its use in support of academic research for mental health interventions. For instance, P5 notes they would support *"some sort of study being done to better understand and assist people mentally, especially since the internet seems to be such a hive of toxic interactions, if it's being used to better understand people's brains or some sort of medical or academic level, I could see where that would be fine."* Overall, some participants had positive associations with and attitudes toward the use of emotional data and development of automatic ER-enabled wellbeing interventions by academic and medical researchers, even when they had more negative attitudes about automatic ER-enabled wellbeing interventions broadly and applied to them individually.

Past work has shown that people are generally either comfortable with or ambivalent about the use of social media data to support academic research in specific contexts, though noted that self-selection bias of participants participating in a research study, and younger, some college educated demographics from their sample from MTurk, an online crowdsourcing marketplace, may have played a role in their findings [82]. Likewise, research has suggested that people are

supportive of the use of social media data to support researchers' population-level monitoring of mental health, even when they generally had privacy concerns; however, this study focused on mostly educated people or people working in a white collar capacity [133]. We surmise that the positionality of our largely college-educated sample, similar to past work with comparable findings [82], may have contributed to these participants' comfort in the use of social media data for research, and may not be generalizable to the broader population.

Participants stressed, however, that automatic ER-enabled wellbeing interventions used to support academic research should only occur if the person has knowledge of the emotion data collection and corresponding intervention. For instance, P5 adds: *"But I wouldn't be okay with that being used without anybody's knowledge because that's just shady. If you're not going to tell people, then why aren't you telling people? I doubt on such a large scale that it would affect their finding."* P7 echoes, *"I would want that information known to me as the user."* Thus, while some participants suggested they might feel comfortable with automatic ER-enabled wellbeing interventions if they were used to support research, they felt that research should be used only in a transparent context where the affected individual is informed of the practice and meaningfully consents to it.

4.2.2 Increasing Access to Wellbeing Support. Several participants suggested that automatic ER-enabled wellbeing interventions could be a positive benefit for those without a strong support network, thereby increasing access to wellbeing support. While P8 was hesitant about the potential of emotion data's use in other ways such as product advertising or targeting them individually, P8 responded positively to its use toward this type of "good." When asked about using emotion data to create wellbeing-related interventions, P8 responded: *"That feels more like the social good side of it, like using this for good rather than like here's a weight loss pill...Like that feels less of a social good than like someone is having an acute moment than like this platform can be used to actually provide resources that might help in that moment."* Here, P8's attitude toward emotion detection on social media platforms is contextual: while P8 expresses discomfort with its use to promote advertising, they respond positively to its use to promote the social good of mental health support. On social media platforms' unique position to offer wellbeing interventions, P6 acknowledges *"there's some people that don't have family to intervene and maybe that would be good for a person who does not have anyone and they're using social media as a cry for help."* P12 echoes *"I think that's good. Again, you see kids on there and they might have a problem. They need help."* People in distress sometimes rely on social media to seek support, and some participants viewed wellbeing interventions as a tool to provide it where users may not otherwise receive it. This acknowledgement echoes a potential motivation for some of the research behind and application of practical platform-based interventions in place today.

It is important to note that while some participants imagined automatic ER-enabled wellbeing interventions as a potential social good that could increase access to wellbeing support, those same participants remained uncomfortable about its use on *themselves*. Future work could consider the impact of social support networks on data subjects' attitudes toward automatic ER-enabled wellbeing interventions.

4.2.3 Egregious Harm Prevention. Perhaps the most widely acknowledged benefit participants imagined automatic ER-enabled wellbeing interventions could offer was its potential to prevent egregious harm, such as in cases of social media users planning suicide, harm toward others, or domestic terrorism. Some participants considered automatic ER-enabled wellbeing interventions to have the potential to be particularly helpful and effective to prevent harm. P11 saw interventions as helpful *"actually for people who are an immediate threat to themselves or others... or mainly themselves"* and remarked, *"I think it could be a very good thing."* P8 worried about the example of *"someone...searching about like ways to commit suicide or ways to hurt someone. I think that's when*

I, I feel like the social good and like someone's bodily safety is at risk," and imagined the potential of interventions to mitigate that risk. For these use cases, participants considered the good of preventing suicide and harm toward others to benefit society at large.

Furthermore, P6 discussed automatic ER-enabled wellbeing interventions as a potential way to reduce the number of school shootings, and thought that *"maybe there would actually be less things occurring"* if there were mental health interventions on social media. P6 elaborated that ER-enabled wellbeing interventions might assist public safety officials by understanding content posted by these individuals: *"we need to monitor posts a little more closely to see if there's somebody who's vaguely talking about a school shooting or something, they say, we need to sometimes be responsive, and not just take someone at their word, because someone's word may not express exactly what they're about to walk out the door and go and do."* This participant's account can be taken to mean that some individuals feel that surveillance on social media can be justified in efforts to prevent domestic terrorism such as school shootings, but that harm can arise when such surveillance leads to inaccurate and biased identification. Considering the rhetoric and justifications used by participants to imagine wellbeing interventions as a positive force that can prevent egregious harm, we speculate that these attitudes may have been influenced by the positive discourse of some high-profile wellbeing interventions such as Facebook's Suicide Prevention Program as tools that promote *wellness* and *help users in need* [5]. Likewise, positive positions may have been shaped by rhetoric in the popular press that AI can help *save the planet* [6] and by the government that AI can *empower people* and *improve peoples' lives* [10].

In summary, some participants conceptualized automatic ER-wellbeing interventions on social media more positively when imagining its greater social impact, going beyond individual concerns. In these instances, participants resolved their tension between their general negative attitudes toward automatic ER-enabled wellbeing interventions and their more positive attitudes when the interventions might benefit others as a possible social good by establishing that automatic ER-enabled wellbeing interventions should be transparent to individuals, and that individuals should meaningfully consent to their use.Future work could delve deeper into such shifts in expectations and attitudes at individual and collective levels.

4.3 Negative Conceptualizations of and Attitudes Toward the Impact of Automatic Emotion Recognition-enabled Wellbeing Interventions for Others

When conceptualizing its impact on others, most participants maintained their negative attitudes toward automatic ER-enabled wellness interventions on social media. These participants were primarily concerned that the intervention might commit harm to other people, and emphasized a need for individual and external control to mitigate those harms.

4.3.1 Potential Harms. Participants expressed a variety of concerns that automatic ER-enabled wellbeing interventions on social media pose a risk of harm to other users, including risks of: re-traumatization, spread of inaccurate health information, inappropriate surveillance, and interventions informed by inaccurate predictions.

Some participants expressed concern about the potential for re-traumatization caused by interventions. For instance, P10 said: *"Like, it could help people. It could also make people more angry that a machine is telling them, 'Hey, you sound angry. Please call this number.' Like, 'All right, machine. Calm down. Leave me alone.'"* In this example we see that some participants feared the interventions themselves may lead to outcomes of anger and frustration to the individual, leading to re-traumatization of already vulnerable individuals.

Others felt that if medical interventions and diagnoses became commonplace on social media, people may start to believe that whatever information they are given about their wellbeing is

accurate and credible. For instance, P7 said: *"It just feels like it's going to put information into the hands of uneducated people who are then going to assume that Facebook is accurate... I feel like it's going to lead to people...overreacting."* In this example, the participant expresses concerns that relying on social media for wellness information can lead to the spread of misinformation, particularly among vulnerable groups such as those with low educational attainment.

Another concern expressed was that the surveillance methods required to enable automatic ER-enabled wellbeing interventions could be applied by individuals in other contexts that could then cause harm through privacy infringement. For example, P3 wondered, *"But again are there parents wanting to use that to monitor their kids? I understand that but I just don't think it would be good to try to...I just feel you'll do more harm than good but that's my fear."* Participants acknowledged that the data collected from constant monitoring could be used and abused by other entities, and were concerned how that might harm certain groups such as children.

Speaking further to potential harms, P6 said: *"I think that maybe there would actually be less things occurring because people use social media now for everything, as I said before, some of the things people post online, I'm like, I can't believe you even put that on there. And maybe it would be very helpful, but at the same time there could be a fine line because what if you're insinuating something else and you end up investigating someone for something that has nothing to do with what you were thinking they were talking about."* Participants worried that automatic ER-enabled wellbeing interventions, especially in cases where the prediction is inaccurate, could harm the intervened subject. Unless the individual had control and agency in the surveillance that facilitates automatic ER-enabled wellbeing interventions, participants felt there would be significant risk that other actors might exploit that surveillance for ethically questionable purposes.

These varied examples show that the harms people imagine automatic ER-enabled wellbeing interventions can commit span a wide range of concerns, and suggest that its potential harm is immense.

4.3.2 Individual and External Control. Overall, participants were concerned about the expression of power in the user-platform relationship when conceptualizing automatic ER-enabled wellbeing interventions and the potential for harm within that context. As P2 put it, *"Assuming that the intervention was not forced intervention, I think it would be a good thing. If the intervention were forced, then I would tend to say things have gone too far."* In this example, we see that people are opposed to any intervention they perceive to be unconsented to and forced upon them. Participants stressed that having the choice to control whether they were subjected to these interventions would allow the intervention to reach the people who might need it, while allowing those more reserved about its outcomes a choice in whether they were subject to the intervention. Participants felt they would be more comfortable with the delivery of automatic ER-enabled wellbeing interventions on social media platforms if there were clearly defined boundaries to help those in need of support, and options to enable and maintain user control. P8 noted the need for bounds and control on the deployment of automatic ER-enabled wellbeing interventions: *"I think about it at an individual level. I don't like that idea. But when I think about [the] crisis that we're in and like I think about queer youth or whomever and things that people are posting about and are like crises that people do post to Facebook around in moments of crises. I think if it helps people who are in that acute moment, then maybe I'm okay with it, but I would want there to be like bounds on that."* P8 was cautious about sanctioning the use of automatic ER-enabled wellbeing interventions for people in crisis, and was sure to underscore the need for measures that would subject the interventions to external regulation and allow for individuals' control before approving of its use.

In these examples we show that participants maintained their negative attitudes toward automatic ER-enabled wellbeing interventions whether they imagined it at a personal or social level.

Participants expressed strong preferences for individuals to have control in whether they were subject to interventions, and for interventions to be subject to external regulation, both of which may mitigate some concerns surrounding potential harms.

4.4 Development and Delivery Qualities Upon Which Attitudes toward Automatic Emotion Recognition-enabled Wellbeing Interventions Depended

While some participants maintained negative attitudes toward automatic ER-enabled wellbeing interventions at all costs, some imagined particular qualities that, if implemented, might engender some increased degrees of comfort and trust in the intervention. We identified three qualities upon which they felt their level of trust and comfort in this technology depended: 1) accuracy; 2) contextual sensitivity; and 3) positive outcome.

4.4.1 Accuracy. Some participants believed that automatic ER-enabled wellbeing interventions for support should be based on highly accurate inferences, and saw potential negative consequences for individual harm should the intervention fail to meet certain expectations of accuracy. For example, P3 said: *"if you see someone caught retweeting about bad shit, and it's like then clearly you should call him if they say they want to die, they want to die. That's not always accurate. So I feel like it would make them completely have that option that, if people are at risk or whatever, for them to use that...but again...I just feel that you'll do more harm than good but that's my fear."* Participants understood that while interventions such as ER-enabled suicide predictions could have potential positive benefits, their accuracy would be a determining factor in whether they helped or harmed the individual user. Participants expected highly accurate algorithms that are able to understand nuanced and contextual engagement with the platform before they would consider themselves comfortable with the automatic ER-enabled wellbeing intervention deployed on social media.

Related to accuracy was the quality of relevance. Some participants expressed a requirement that they perceive automatic ER-enabled wellbeing interventions as relevant to their condition. These participants might feel more comfortable with the idea of automatic ER-enabled wellbeing interventions, so long as those interventions were accurate enough to offer relevant support to them. For instance, P12 said: *"[Y]ou might be able to learn something about yourself and about the condition too. I think it's great, it's free help, you know? As long as it's a credible source...you can learn a lot about new treatments, and therapy, and that type of thing. It might even help you because maybe you've tried all these different medicines and remedies, and you're not getting anywhere. Now they have a new breakthrough, wow look at this. I'm always researching, and always looking into new things. I would like that. It might be really good, it might help me."* Participants imagined that interventions that were accurate enough to have specific relevance to their individual conditions could then be helpful, through the advancement of individual knowledge about the relevant condition and its treatment options.

These examples highlight the importance of accurate automatic ER-enabled wellbeing interventions, yet suggest that they should be optional for the data subjects (not all desire precise accuracy, and some just desire relevance), provide customized support, and be relevant to their condition.

4.4.2 Contextual Sensitivity. The specific wellbeing context in which interventions were provided mattered to some participants. For example, an imagined intervention suggesting resources in one's geographic area about a physical illness was seen to be less intrusive than resources regarding mental illness. To this point, P7 said: *"Let's say I have some rare medical condition and it shows me an ad for a clinical trial in my area, that could save my life. But yeah, I don't know why, maybe it's such a stigma, but for some reason if it's a mental health thing, that seems more slimy to me that they're advertising towards that, that they're taking advantage of me. But if it's like any other health issue it doesn't seem as slimy."* Here we see that some participants felt that interventions for physical health

conditions might be helpful, but felt that mental health related interventions were too intrusive and exploitative. P8 commented, as discussed in 4.2.3, that only in specific contexts, such as preventing harm toward others or themselves, that automatic ER-enabled wellbeing interventions could be a positive benefit to the community. P8 explains: *"If someone were searching about like ways to commit suicide or ways to hurt someone. I think that's when I, I feel like the social good and like someone's bodily safety is at risk, you know, theirs or someone else's. It feels like that's a time when the fact that this is all one soup, that should be used, but I think that would probably be the line for me."* In this example, we see that people with overall negative attitudes toward automatic ER-enabled wellbeing interventions might temper their objection in contexts where the technology's potential for what they deemed as social good outweighs their own reservations.

Our results indicate that participants' comfort level with the deployment of automatic ER-enabled wellbeing interventions was highly dependent on the context in which the intervention would be used. For what types of interventions people would welcome automatic ER-enabled wellbeing interventions is an area for future research, but is certainly not a trivial question. What is more, our findings show that assuming that all automatic ER-enabled wellbeing-related interventions would be welcomed by data subjects is inappropriate.

4.4.3 Positive Outcome. Some participants' attitudes were dependent upon tangible impacts the intervention may have on them. In 4.3.1 we describe how anticipating harm was a reason for negatively held attitudes toward automatic ER-enabled wellbeing interventions. Here we describe how if the intervention were proven helpful, participants might be more comfortable with it; if the outcome were not helpful, they would not welcome the intervention. For instance, P7 said: *"Because if it's successful and I feel better, then I feel like I can't be upset about it."* P10 echoed similar attitudes, and additionally suggested that a layer of assurance such as a certification process would increase their confidence in the positive outcome: *"I think that I would feel okay with that, as long as that support is I guess somehow certified or goes through a process of guaranteeing that it's not shitty so I feel worse. I think I could support that use of data."* Participants felt that if the outcome of the intervention were successful, then they could embrace its use. Our findings show that data subjects have strong preferences for automatic ER-enabled wellbeing interventions to assure *positive* outcomes on them.

In summary, these insights into development and delivery qualities of automatic ER-enabled wellbeing interventions upon which data subjects' attitudes depend suggest some individuals may welcome accurate, contextually sensitive ER-enabled wellbeing interventions with guaranteed positive outcomes. Who, to what extent, and in what contexts, would welcome interventions developed and implemented with such preferences is an area for future work.

5 DISCUSSION

Our study examined data subjects' conceptualizations of and attitudes toward automatic ER-enabled wellbeing interventions on social media. At a high level, we contribute to discourse around the development of socially aware, trustworthy, and ethically responsible AI advancements, with a focus on emotion-sensitive technologies. Specifically, we contribute a characterization of data subjects': 1) broad conceptualizations of automatic ER-enabled wellbeing interventions; 2) positive conceptualizations of and attitudes toward the impact of automatic ER-enabled wellbeing interventions for others; 3) negative conceptualizations of and attitudes toward the impact of automatic ER-enabled wellbeing interventions for others; and 4) development and delivery qualities upon which their attitudes toward automatic ER-enabled wellbeing interventions depend.

We suggest that data subjects' negative conceptualizations of automatic ER-enabled wellbeing interventions are shaped by a human versus AI dichotomy and beliefs that automatic ER-enabled wellbeing interventions could not hold attributes of supportive wellbeing interventions traditionally

delivered by humans: 1) helpfulness and authentic care; 2) personal and professional expertise; 3) morality; 4) benevolence through shared humanity.

Our findings reveal the potential for relatively more positive conceptualizations of automatic ER-enabled wellbeing interventions to show their presence when imagining the social impact of automatic ER-enabled wellbeing interventions on others. Some imagined the tool as a potential social good that could benefit others: 1) by supporting academic research; 2) by increasing access to wellbeing support, and 3) through egregious harm prevention. These positive attitudes are complicated by participants' concerns of potential harms that automatic ER-enabled wellbeing interventions could present to others (e.g., re-traumatization, spread of inaccurate health information, inappropriate surveillance, and interventions informed by inaccurate predictions). Even when imagining interventions as a social good, participants expressed requirements that automatic ER-enabled wellbeing interventions are transparent to individuals, that individuals meaningfully consent to them, that individuals have control over their use, and that interventions are subject to external regulation.

Lastly, we contribute a characterization of development and delivery qualities upon which data subjects' attitudes toward automatic ER-enabled wellbeing interventions depended: 1) accuracy; 2) contextual sensitivity; 3) positive outcome. As such, we identify what makes (and does not make) for an ethical and trustworthy automatic ER-enabled wellbeing intervention on social media.

While we found that data subjects' attitudes track well to similar themes of harm and privacy concerns found in the literature critical of ER [33, 48, 53, 89, 125, 157], our study builds on recent work [22, 85] and *empirically* centers the voices and concerns of the humans that make the technology possible to begin with — and those subject to its consequences — rather than merely approaching this discourse from an abstract perspective.

We align ourselves with human-centered computing (HCC) [107] and social constructivism [154] approaches and include *humans* as relevant social groups [154] and stakeholders in our study to contribute to requirements and considerations for ethical and trustworthy ER applications. Our goal is not to make normative statements about whether automatic ER-enabled mental health interventions *should* exist, but rather to complicate existing discourse surrounding this technology through promoting the voices and concerns of the humans most impacted by it.

5.1 What Makes an Ethical and Trustworthy Automatic Emotion Recognition-enabled Wellbeing Intervention on Social Media?

Participants in our study were overall consistent and clear in their rejection of automatic ER-enabled wellbeing interventions on social media: they neither wanted nor needed it, including those who spoke from personal mental health experiences. Participants did not trust automatic ER-enabled wellbeing interventions on social media to deliver support in the way humans can, and were concerned about the potential harm interventions could cause others, including re-traumatization, spread of inaccurate health information, inappropriate surveillance, and interventions informed by inaccurate predictions. Compared to human support, they deemed automatic ER-enabled wellbeing interventions as unhelpful, immoral, incompetent, and ineffectual. Even for those few participants that held slightly more positive attitudes regarding automatic ER-enabled wellbeing interventions when conceptualizing its impact on others, the benefits they imagined were counterbalanced by concerns including potential harm to individuals. These insights reflect past findings by commercial research and advisory firm Gartner, that showed out of 4,000 US and UK respondents, more than 52 percent did not want their faces to be subject to affect recognition [128]. While participants expressed requirements and qualities that might improve their attitudes toward or trust in automatic ER-enabled wellbeing interventions, these requirements and qualities are incompatible with current social media practices, and might be challenging to deliver (i.e., a guaranteed positive outcome).

Our findings complement work that centers human perspectives in understandings of wellbeing interventions in other contexts. For example, past work has shown that older adults express a willingness to use smart home technologies in support of self-management of wellbeing [67]; however, has not focused specifically on emotions. Yet emotions are a sensitive and unique kind of data, different from other types of data that people may deem private [22]. Future work is needed to identify the contexts in which people may welcome automatic ER-enabled wellbeing interventions with more nuance. For example, it may seek to understand how older adults perceive of the use of automatic ER-enabled wellbeing interventions using voice assistants, rather than wellbeing interventions delivered via smart home technologies broadly. Our work provides preliminary insights that data subjects are hesitant to receive automatic ER-enabled wellbeing interventions on social media. In addition, our paper's findings identifying negative conceptualizations of and attitudes toward the impact of automatic ER-enabled wellbeing interventions for others resonate with past work on human-AI collaboration, showing that "trust is the most correlated with human preferences of optimal human-machine delegation" [120] and that without trust, humans are not likely to feel comfortable with the delegation of traditionally human tasks to AI [108] (and as we find, especially not those as intimate as wellbeing).

If data subjects neither want nor need automatic ER-enabled wellbeing interventions on social media, socially aware and ethically responsible design must listen. People should not be subject to such an opaque and invasive technology through which social media companies capitalize human emotion, and consequently present harm to its data subjects. More work is needed to identify in what contexts people might welcome automatic ER-enabled wellbeing interventions, such as in a non-commercial medical context under the supervision of medical providers — and by proxy, medical data privacy protection and regulations. Our work has shown that data subjects have overall negative attitudes toward automatic ER-enabled wellbeing interventions in the context of social media, and have clear and specific requirements for accuracy, contextual sensitivity, and positive outcome before they could welcome such interventions on social media. Based on our findings, we urge social media platforms that have deployed (or are considering deploying) automatic ER-enabled wellbeing interventions to align their applications with data subjects' requirements for trustworthy delivery of automatic ER-enabled wellbeing interventions on social media.

5.2 But What if Individuals Consent?

We acknowledge that despite data subjects voicing alternative preferences, automatic ER-enabled wellbeing interventions and ER more broadly will continue to expand. ER is projected to be a twenty-five billion dollar market by 2023 [12], and has current applications in industries that impact the lives of the population at large, including law enforcement [121], recruitment [46], financial services [84], medicine [98], education [1], and advertising [110]. In practice, many people are subject to ER without either their knowledge or consent, and ER's commercial viability and growth suggests that this trend will continue. For example, the Chromebooks used by children in over ten thousand schools across North America are subject to an educational management and monitoring system, GoGuardian. Its Beacon module, an automatic ER-enabled suicide prevention and early detection tool, is offered to all of GoGuardian's admin customers at no additional cost [13]. Beacon algorithmically monitors "web searches, social media, chat, forums, email, and online collaboration tools" to detect students' mental state and predict violence and safety threats, under a veneer of *safety through surveillance* [4]. Children and their parents have little to no option to opt-out, as GoGuardian "obtains school-based consent under the Children's Online Privacy and Protection Act (COPPA)" [11].

In another far-reaching example, Facebook's suicide intervention program scans all posts on the social media site for risk of imminent harm, with no option for individuals to opt out. In response to

a journalist's inquiry, a Facebook representative explained: "By using Facebook, you are opting into having your posts, comments, and videos (including FB live) scanned for possible suicide risk" [32]. Facebook's suggestion that its users consent to all of their data practices by using Facebook is rooted in the "notice and choice" framework the Federal Trade Commission (FTC) uses to safeguard data privacy. Under this model, online information providers (and collectors) are required to disclose to consumers their data practices, and then the consumer decides whether or not to continue with the service [139].

Notice and choice is the industry standard for privacy policies, offering online consumers a restrictive, binary option: accept the provider's terms in order to use the product, or opt-out entirely. The decision of whether to click 'I agree' has much more at stake than use of the platform itself. As several privacy scholars have argued, notice and choice presents a false choice for consumers: we live in an increasingly connected world in which engagement with online platforms becomes increasingly necessary to engage with the modern world [41]. Social media platforms play an important role in the way humans use, create, and maintain social capital [73]. Marginalized communities in particular depend on online information and social networks to seek support and community (e.g., [20, 23, 26, 28, 93]).

While social media plays a crucial role in humans' social capital, information access, and wellbeing, platforms themselves rely upon the commodification of the personal data people produce to sustain their business model [100]. Despite public calls for greater individual control and agency over the use and sharing of personal data on social media [101] — calls echoed by the participants in our study — platforms flex their strong position in the power asymmetry between social media platform and data subject by failing to implement tailorable and context-sensitive privacy controls [30, 137]. Instead, they offer only the binary option to accept their terms of service entirely or opt out of their service entirely. For those that try to read them [88], privacy notices are written in often obtuse, hard to understand language heavily slanted toward the interests of the service provider, with little regard for consumer interests [81, 86, 118]. Opting out of such sites as social media presents an enormous social and personal cost to individuals. To the already marginalized people that rely upon social media for crucial information and support, forcing a choice between information access and community, or privacy, autonomy, and control, only further disadvantages them while sustaining the power imbalance between data subjects and the corporations that collect and commodify their data, livelihoods, and experiences.

Thus, platforms that fall back on the traditional "notice and choice" argument in data collection (including automatic ER-enabled wellbeing interventions) and fail to take these criticisms into account when employing invasive, controversial technology are at odds with advances to promote ethical and socially responsible AI technologies. In current practice, inferences about mental health data are made on unwitting individuals with little to no regulatory oversight over the collection, protection, and dissemination of those inferences, under the pretense of protecting a small fraction of individuals.

More work is needed to explore alternatives to the "notice and choice" framework, and how people might actually welcome and benefit from AI-driven interventions, and not simply get accustomed to them as Zuboff warns [169], particularly in the context of platforms that employ automatic ER-enabled wellbeing interventions. Our findings indicating that some data subjects acknowledge automatic ER-enabled wellbeing interventions on social media as a potential social good, yet are 1) concerned about its potential harms, 2) desire individual and external controls in its application, and 3) qualify that such interventions should be accurate, contextually-sensitive, and guarantee positive outcomes to the data subject, provide fertile groundwork for this important future work.

We draw attention to our finding that some people felt more positively about automatic ER-enabled wellbeing interventions if they were developed in concert with academic researchers. As Google ethicist Alex Hanna and AI Now Institute co-founder Meredith Whittaker have recently pointed out, the corporate gatekeepers of AI enjoy a close relationship with academic researchers by providing significant funding to top computer science departments, offering concurrent positions to researchers who hold appointments at universities, and publishing papers together. "This blurs the boundary between academic and corporate research and obscures the [economic] incentives underwriting such work" [96]. Highlighting the case of Google's recent act of firing Timnit Gebru — co-lead of Google's ethical AI team who researches racial and gender bias in AI systems and was let go after Google demanded she rescind a paper under peer review that exposed bias in (highly profitable) large language models — Hanna and Whittaker warn that "powerful companies like Google have the ability to co-opt, minimize, or silence criticisms of their own large-scale AI systems—systems that are at the core of their profit motives" [96]. We caution that collaborations between social media platforms and academic researchers developing automatic ER-enabled wellbeing interventions on their platforms might obviate data subjects' requirements for its development and delivery qualities of accuracy, contextual sensitivity, and positive outcome, by manipulating peoples' trust in academic institutions to silence criticism.

5.3 Harm to Vulnerable Populations

Emotion data should be considered sensitive in research and practice [22]. While automatic ER-enabled wellbeing interventions can target any individual whose emotions can be inferred or predicted from their online behavior (our focus), ER's harms might be most acutely felt by certain vulnerable populations. In a healthcare context, vulnerable populations are defined as those "at greater risk for poor health status and healthcare access" and include the economically disadvantaged, racial and ethnic minorities, and those with chronic health conditions including mental illness, with vulnerability increasing with factors such as "race, ethnicity, age, sex, and factors such as income, insurance coverage...and absence of a usual source of care" [2]. Mental health patients are an exceptionally vulnerable population in the unregulated space of ER and wellbeing interventions: they are subject to involuntary, coerced care more than any other population [160], potentially exacerbated by unregulated intervention programs. Recent work exploring mental health related apps and digital phenotyping involving technology companies broadly has shown that individuals with mental illness are wary of algorithmic inferences made of health status and associated advertising from their use, and echo many of the concerns with mental health applications and mental health condition inferences that our study's participants had regarding corporate profit motives, distrust, and calls for controls such as external regulation [57]. As our findings show, data subjects are also concerned that automatic ER-enabled wellbeing interventions on social media carry significant risk of harms such as re-traumatization. For those living with mental illnesses that seek support on social media, their use of the platform might result in unwanted (and unwarranted) traumatic experiences.

Inferences made regarding mental health states can hold grave consequences, especially for racial and ethnic minorities that have been shown to be more likely to be admitted involuntarily to mental health institutions [129]. Further, the interventions that rely on those inferences, such as Facebook's suicide intervention which surrenders an individual's personal information to police, who then respond with a 'welfare check,' when it infers an individual is in need of crisis, may subject certain communities to adverse harm. Police encounters between people in behavioral crisis and police often end in unwarranted brutality [146], an outcome already disproportionately affecting Black, Brown, and Indigenous people in the US [112, 152]. When the police are called to respond to the mental health crisis of a person of color, it is an all too recurrent outcome that the

individual in crisis will not only receive inadequate care, but will be subjected to police violence instead [38, 162]. In addition to a concern for harms such as re-traumatization, participants in our study expressed a concern for harm to data subjects from interventions based upon inaccurate ER inferences. The algorithm's false positives could present harm from law enforcement involvement when a person was never in crisis in the first place, leading to uncalled for and unjustified risk. Future work should examine the impact of automatic ER-enabled wellbeing interventions that include protocols to involve police for mental health calls on individuals with experience being targeted by them.

Recent research has shown the feasibility of detecting emotion and "violence estimation" from social media data [165], work in which the US government has shown interest in deploying [36]. In light of the civil unrest and cultural reckoning the US has experienced with the revival of the Black Lives Matter movement in 2020, these predictions of protest activity — and their co-predictions of violent risk — from social media raise questions about the role of data harvesters and their responsibilities to the individuals that enable their technology. The dissemination of social media data that can be and has been used to target a population already disproportionately criminalized [112, 152] might produce chilling effects in civil rights protest activity, reifying and perpetuating white supremacist power structures. We urge social media companies to consider ways to prevent these alarming uses of social media data, such as a screening measure when sharing data with third parties [106].

Future work is needed to understand the preferences and needs of diverse communities made vulnerable regarding automatic ER-enabled wellbeing interventions. We urge social media platforms to thoroughly consider how existing wellbeing interventions can eliminate their harm to (and even protect and benefit) data subjects who would be at most risk.

5.4 The Tension between Monitoring for Harm Prevention and Individual Privacy

Empirical work has suggested that the apparent contradiction between individual actions in loosely sharing and disclosing information online and strong individual preferences for privacy can be resolved when understanding the nuanced contextual variables in which people disseminate information [126]. For example, sharing sensitive information such as health data within commercial flows (i.e., with a health insurance agency, or at a doctor's office) generally meets privacy expectations within that particular, appropriate context, but the subsequent sharing of that same information in another context — say, to one's employer or made available to public record — generally does not meet peoples' privacy expectations [126]. Our study found that while people held generally negative attitudes toward automatic ER-enabled wellbeing interventions, some participants adopted a positive attitude when imagining its use in limited use cases, such as to prevent egregious harm. However, the methods required to employ an intervention tool that prevents harm necessarily means that individuals cannot be granted their preferences for privacy of emotional inferences or to share that information in contextual, nuanced settings: the algorithms must scan most or all content to be effective, thus violating the contextual integrity of the disclosed information [141, 155]. The participants in our study stressed a preference for individual autonomy and control over being subject to automatic ER-enabled wellbeing interventions on social media, a design option that would enable individuals to control the sharing of information they disclose online.

We argue that rather than designing privacy controls that respect individual preferences to control sensitive information sharing, which would restrict social media platforms' commodification of valuable user data, platforms instead have focused discursive efforts to influence social norms such as those viewing interventions as a tool that promote public safety. For example, Facebook has framed their Suicide Prevention algorithm as an *AI-fueled detection effort* that provides *timely help*

to *people in need* [5]. GoGuardian, which contracts with school districts to monitor student devices, has promoted its AI-enabled behavioral risk detection as a tool that promotes student *safety* and identifies students *in need* of a *psychological intervention* [127]. We speculate, based on our findings, that discursive strategies to frame automatic ER-enabled wellbeing interventions as a way to promote public safety have likely worked: participants in our study who reported to generally feel negatively toward automatic ER-enabled wellbeing interventions targeting themselves, somewhat contradictorily, felt they might positively impact society by preventing egregious harm. We suggest this tension might be explained by the influence of public relations efforts pushed by companies that have deployed wellbeing interventions to frame them as a positive social good, and discourse in general by the popular press and government framing AI as a human savior [6, 10]. These efforts, we suggest, gently shift social norms of mass surveillance toward acceptance [86].

As Shoshana Zuboff has argued, surveillance capitalists (as well as governments) have a vested interest in nudging people to abandon privacy and accept data collection, a practice from which surveillance capitalists financially and strategically benefit [169]. Indeed, Facebook's CEO Mark Zuckerberg has famously and controversially said that "privacy is no longer a social norm" [111]. This suggestion aligns with past work that has argued that people "naively or unwittingly trust their personal information to corporate platforms" and extend that trust to data-sharing with external parties such as law enforcement [66]. We surmise that the discourse by powerful actors painting AI as a tool for human salvation, along with the trust people generally place in corporate platforms, has contributed to the approval of some and apathetic acceptance of others to automatic ER-enabled wellbeing technologies. We urge caution of these corporate strategies to promote unfounded acceptance of and trust in mass monitoring, especially of emotions, masquerading as a public good.

6 CONCLUSION

Through centering data subjects' conceptualizations of and attitudes toward automatic ER-enabled wellbeing interventions on social media, we contribute to discourse around the development of socially aware, trustworthy, and ethically responsible AI advancements. We found that people have predominantly negative attitudes toward automatic ER-enabled wellbeing interventions, and conceptualize harmful consequences including re-traumatization, spread of inaccurate health information, inappropriate surveillance, and inaccurate predictions. We find that data subjects' attitudes toward automatic ER-enabled wellbeing interventions were rooted in their conceptualizations of the human versus AI dichotomy, and human attributes they doubted wellbeing interventions could hold. We also found that people conceptualize different concerns when thinking of the impact of automatic ER-enabled wellbeing interventions for others, rather than at a general or personal level. We identified qualities in either the development or delivery of the intervention upon which attitudes depended. We argue that technology companies that deliver or consider delivering automatic ER-enabled wellbeing interventions ought to consider the attitudes and concerns of the data subjects that enable their technology — and those vulnerable to its potential harms — in alignment with proposed industry goals to promote ethical and socially aware AI applications. Participants in our study (including those with real mental health-related experiences) did not want to be subjected to automatic ER-enabled wellbeing interventions and had difficulty imagining a need for them. Imposing people to such exploitative technology when they neither want nor need it — and when they do not have explicit knowledge about it — is nontransparent and ethically questionable. We argue that to increase the trustworthiness of automatic ER-enabled wellbeing interventions on social media, companies that deploy them would need to *at least* fulfill requirements that preemptively protect individuals from the vast harms it presents, take measures to attenuate harms, and align with data subjects' development and design requirements. These requirements include high

computational accuracy, contextual sensitivity, positive outcome guarantees, individual controls, external regulation, and meaningful consent over being subject to automatic ER-enabled wellbeing interventions. We conclude with a message of caution and restraint about the use of automatic ER-enabled wellbeing interventions on social media in the US, based on its current regulatory landscape and social context.

ACKNOWLEDGMENTS

We are thankful to the participants who generously shared their perspectives with us, to Justin Buss for assisting in collecting the data, to Cassidy Pyle and Dan Delmonaco for comments on earlier drafts, to Emily Sartorius for earlier support with related work, and to the anonymous reviewers and associate chairs for their constructive and encouraging feedback. We are also thankful to the National Science Foundation (award number 2020872) for supporting this work.

REFERENCES

[1] [n.d.]. A Suicide Prevention Tool for Schools | GoGuardian Beacon. https://www.goguardian.com/beacon/
[2] 2006. Vulnerable Populations: Who Are They? https://www.ajmc.com/view/nov06-2390ps348-s352
[3] 2018. Gartner Projections for 2018. *Database and Network Journal* 48, 2 (April 2018), 10–. http://link.gale.com/apps/doc/A539646974/AONE?u=umuser&sid=zotero&xid=a4a6464b Section: 10.
[4] 2018. GoGuardian Launches Beacon Tool. *Health & Beauty Close-Up* (Aug. 2018). https://bit.ly/33VEFjt Publisher: Close-Up Media, Inc.
[5] 2018. How Facebook AI Helps Suicide Prevention. https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/ Section: Facebook.
[6] 2019. AI could be a critical tool to help save the planet. *The Guardian* (April 2019). https://www.theguardian.com/ai-for-earth/2019/apr/30/ai-tech-sustainable-planet
[7] 2019. Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences, 5th Edition. *ProtoView* 2019, 31 (Jan. 2019). https://bit.ly/34XakAi Place: Beaverton Publisher: Ringgold, Inc.
[8] 2019. U.S. Census Bureau QuickFacts: Detroit city, Michigan; Michigan. https://www.census.gov/quickfacts/fact/table/detroitcitymichigan,MI/PST045219
[9] 2019. U.S. Census Bureau QuickFacts: Houston city, Texas. https://www.census.gov/quickfacts/houstoncitytexas
[10] 2020. Artificial Intelligence for the American People. https://trumpwhitehouse.archives.gov/ai/
[11] 2020. Communicating with Parents/Guardians. http://help.goguardian.com/hc/en-us/articles/360025096772
[12] 2020. Emotion Analytics Market 2020 - Recent Development and its impact on Market Share, Size, Sale, Growth Rate and Future Opportunity. https://www.marketwatch.com/press-release/emotion-analytics-market-2020---recent-development-and-its-impact-on-market-share-size-sale-growth-rate-and-future-opportunity-2020-09-16
[13] 2020. GoGuardian Offers Suicide Alert Software to All of Its Admin Customers for Free. *Entertainment Close-up* (Jan. 2020). http://link.gale.com/apps/doc/A611723350/BIC?u=umuser&sid=zotero&xid=19e7b96e Section: NA.
[14] Regulation (EU) 2016/679. 2016. Data Protection Act 2018, c. 12. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN
[15] Hessa Albalooshi, Shahram Rahmanian, and Rahul Venkatesh Kumar. 2018. EmotionX-SmartDubai_NLP: Detecting User Emotions In Social Media Text. In *SocialNLP@ACL*.
[16] Talayeh Aledavood, Ana Maria Triana Hoyos, Tuomas Alakörkkö, Kimmo Kaski, Jari Saramäki, Erkki Isometsä, and Richard K. Darst. 2017. Data Collection for Mental Health Studies Through Digital Platforms: Requirements and Design of a Prototype. *JMIR research protocols* 6, 6 (June 2017), e110. https://doi.org/10.2196/resprot.6919
[17] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Transactions on Computer-Human Interaction* 26, 3 (April 2019), 17:1–17:28. https://doi.org/10.1145/3311956
[18] Tawfiq Ammari, Meredith Ringel Morris, and Sarita Yardi Schoenebeck. [n.d.]. Accessing Social Support and Overcoming Judgment on Social Media among Parents of Children with Special Needs. ([n. d.]), 10.
[19] Nazanin Andalibi. 2017. Self-disclosure and Response Behaviors in Socially Stigmatized Contexts on Social Media: The Case of Miscarriage. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 248–253. https://doi.org/10.1145/3027063.3027137
[20] Nazanin Andalibi. 2019. What happens after disclosing stigmatized experiences on identified social media: Individual, dyadic, and social/network outcomes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing*

Systems. 1–15.

[21] Nazanin Andalibi. 2020. Disclosure, Privacy, and Stigma on Social Media: Examining Non-Disclosure of Distressing Experiences. *ACM Trans. Comput.-Hum. Interact.* 27, 3, Article 18 (May 2020), 43 pages. https://doi.org/10.1145/3386600

[22] Nazanin Andalibi and Justin Buss. 2020. The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376680

[23] Nazanin Andalibi and Andrea Forte. 2018. Announcing Pregnancy Loss on Facebook: A Decision-Making Framework for Stigmatized Disclosures on Identified Social Network Sites. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, Montreal QC, Canada, 1–14. https://doi.org/10.1145/3173574.3173732

[24] Nazanin Andalibi and Andrea Forte. 2018. Responding to Sensitive Disclosures on Social Media: A Decision-Making Framework. *ACM Transactions on Computer-Human Interaction* 25, 6 (Dec. 2018), 31:1–31:29. https://doi.org/10.1145/3241044

[25] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. 2018. Social Support, Reciprocity, and Anonymity in Responses to Sexual Abuse Disclosures on Social Media. *ACM Transactions on Computer-Human Interaction* 25, 5 (Oct. 2018), 28:1–28:35. https://doi.org/10.1145/3234942

[26] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 3906–3918.

[27] Nazanin Andalibi, Margaret E. Morris, and Andrea Forte. 2018. Testing Waters, Sending Clues: Indirect Disclosures of Socially Stigmatized Experiences on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 19:1–19:23. https://doi.org/10.1145/3274288

[28] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: the case of# depression. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1485–1500.

[29] Florian Arendt, Mario Haim, and Sebastian Scherr. 2020. Investigating Google's suicide-prevention efforts in celebrity suicides using agent-based testing: A cross-national study in four European countries. *Social Science & Medicine* (Feb. 2020), 112692. https://doi.org/10.1016/j.socscimed.2019.112692

[30] Louise Barkhuus. 2012. The mismeasurement of privacy: using contextual integrity to reconsider privacy in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 367–376.

[31] Ian Barnett and John Torous. 2019. Ethics, Transparency, and Public Health at the Intersection of Innovation and Facebook's Suicide Prevention Efforts. *Annals of Internal Medicine* 170, 8 (Feb. 2019), 565–566. https://doi.org/10.7326/M19-0366 Publisher: American College of Physicians.

[32] Benjamin Goggin. 2019. Inside Facebook's suicide algorithm: Here's how the company uses artificial intelligence to predict your mental state from your posts. *Business Insider* (June 2019). https://www.businessinsider.com/facebook-is-using-ai-to-try-to-predict-if-youre-suicidal-2018-12 Journal Abbreviation: Business Insider Publisher: Insider, Inc.

[33] Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical Research Protocols for Social Media Health Research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 94–102. https://doi.org/10.18653/v1/W17-1612

[34] Alan L. Berman and Gregory Carter. [n.d.]. Technological Advances and the Future of Suicide Prevention: Ethical, Legal, and Empirical Challenges. *Suicide and Life-Threatening Behavior* n/a, n/a ([n. d.]). https://doi.org/10.1111/sltb.12610 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/sltb.12610.

[35] Dinesh Bhugra, Allan Tasman, Soumitra Pathare, Stefan Priebe, Shubulade Smith, John Torous, Melissa R Arbuckle, Alex Langford, Renato D Alarcón, Helen Fung Kum Chiu, Michael B First, Jerald Kay, Charlene Sunkel, Anita Thapar, Pichet Udomratn, Florence K Baingana, Dévora Kestel, Roger Man Kin Ng, Anita Patel, Livia De Picker, Kwame Julius McKenzie, Driss Moussaoui, Matt Muijen, Peter Bartlett, Sophie Davison, Tim Exworthy, Nasser Loza, Diana Rose, Julio Torales, Mark Brown, Helen Christensen, Joseph Firth, Matcheri Keshavan, Ang Li, Jukka-Pekka Onnela, Til Wykes, Hussien Elkholy, Gurvinder Kalra, Kate F Lovett, Michael J Travis, and Antonio Ventriglio. 2017. The WPA- Lancet Psychiatry Commission on the Future of Psychiatry. *The Lancet Psychiatry* 4, 10 (Oct. 2017), 775–818. https://doi.org/10.1016/S2215-0366(17)30333-4

[36] Sam Biddle. 2020. Police Surveilled George Floyd Protests With Help From Twitter-Affiliated Startup Dataminr. https://theintercept.com/2020/07/09/twitter-dataminr-police-spy-surveillance-black-lives-matter-protests/

[37] Michael L. Birnbaum, Sindhu Kiranmai Ernala, Asra F. Rizvi, Munmun De Choudhury, and John M. Kane. 2017. A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. *Journal of Medical Internet Research* 19, 8 (2017), e289. https://doi.org/10.2196/jmir.7956

[38] Jacob Bor, Atheendar S Venkataramani, David R Williams, and Alexander C Tsai. 2018. Police killings and their spillover effects on the mental health of black Americans: a population-based, quasi-experimental study. *The Lancet*

392, 10144 (2018), 302–310.
[39] Gordon H. Bower. 1981. Mood and memory. *American Psychologist* 36, 2 (1981), 129–148. https://doi.org/10.1037/0003-066X.36.2.129 Place: US Publisher: American Psychological Association.
[40] Jed R. Brubaker, Lynn S. Dombrowski, Anita M. Gilbert, Nafiri Kusumakaulika, and Gillian R. Hayes. 2014. Stewarding a legacy: responsibilities and relationships in the management of post-mortem data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 4157–4166. https://doi.org/10.1145/2556288.2557059
[41] Finn Brunton and Helen Nissenbaum. 2015. *Obfuscation: a user's guide for privacy and protest.* The MIT Press, Cambridge, Massachusetts London, England. OCLC: 927953450.
[42] M. Calvo and L. Nummenmaa. 2007. Processing of unattended emotional visual scenes. *Journal of experimental psychology. General* (2007). https://doi.org/10.1037/0096-3445.136.3.347
[43] Rafael A. Calvo and Sidney D'Mello. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing* 1, 1 (Jan. 2010), 18–37. https://doi.org/10.1109/T-AFFC.2010.1 Conference Name: IEEE Transactions on Affective Computing.
[44] J. M. Carroll. 2000. Five reasons for scenario-based design. *Interacting with Computers* 13, 1 (Sept. 2000), 43–60. https://doi.org/10.1016/S0953-5438(00)00023-0 Publisher: Oxford Academic.
[45] Alessia Celeghin, Matteo Diano, Arianna Bagnis, Marco Viola, and Marco Tamietto. 2017. Basic Emotions in Human Neuroscience: Neuroimaging and Beyond. *Frontiers in Psychology* 8 (Aug. 2017). https://doi.org/10.3389/fpsyg.2017.01432
[46] Electronic Privacy Information Center. [n.d.]. EPIC - In re HireVue. https://epic.org/privacy/ftc/hirevue/
[47] Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. 2019. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 147:1–147:32. https://doi.org/10.1145/3359249
[48] Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. ACM Press, Atlanta, GA, USA, 79–88. https://doi.org/10.1145/3287560.3287587
[49] Stevie Chancellor, Yannis Kalantidis, Jessica A. Pater, Munmun De Choudhury, and David A. Shamma. 2017. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3213–3226. https://doi.org/10.1145/3025453.3025985
[50] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 1201–1213.
[51] Helen Christensen, Philip J. Batterham, and Bridianne O'Dea. 2014. E-Health Interventions for Suicide Prevention. *International Journal of Environmental Research and Public Health* 11, 8 (Aug. 2014), 8193–8212. https://doi.org/10.3390/ijerph110808193 Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.
[52] Sven-Åke. Christianson. 1992. *The Handbook of emotion and memory : research and theory.* Number xix, 507 p. :. Erlbaum Associates, Hillsdale, N.J. :. Publication Title: The Handbook of emotion and memory : research and theory.
[53] Mike Conway and Daniel O'Connor. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current Opinion in Psychology* 9 (June 2016), 77–82. https://doi.org/10.1016/j.copsyc.2016.01.004
[54] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Baltimore, Maryland, USA, 51–60. https://doi.org/10.3115/v1/W14-3207
[55] Glen A Coppersmith, Craig T Harman, and Mark H Dredze. [n.d.]. Measuring Post Traumatic Stress Disorder in Twitter. ([n. d.]), 4.
[56] Juliet Corbin and Anselm Strauss. 2008. *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory.* SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States. https://doi.org/10.4135/9781452230153
[57] Kaitlin L Costello and Diana Floegel. 2020. "Predictive ads are not doctors": Mental health tracking and technology companies. *Proceedings of the Association for Information Science and Technology* 57, 1 (2020), e250.
[58] Nick Couldry and Ulises Ali Mejias. 2019. *The costs of connection: how data is colonizing human life and appropriating it for capitalism.* Stanford University Press, Stanford, California.
[59] Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, Deborah Raji, Joy Lisi Rankin, Rashida Richardson, Jason Schultz, Sarah Myers West, and Meredith Whittaker. [n.d.]. *AI Now 2019 Report.* Technical Report. 100 pages.

[60] Charles Darwin, 1809-1882. 1872. *The expression of the emotions in man and animals.* Number vi, 374, 4 p., 7 leaves of plates (3 fold.) (4 p. at end advertisements) :. J. Murray, London :. Publication Title: The expression of the emotions in man and animals.

[61] Deborah Sarah. David and Robert. Brannon. 1976. *The Forty-nine percent majority : the male sex role.* Number xiv, 338 p. ;. Addison-Wesley Pub. Co., Reading, Mass. :. Publication Title: The Forty-nine percent majority : the male sex role.

[62] Munmun De Choudhury, Scott Counts, and Michael Gamon. [n.d.]. Not All Moods are Created Equal! Exploring Human Emotional States in Social Media. ([n. d.]), 8.

[63] Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 626–638. https://doi.org/10.1145/2531602.2531675

[64] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. AAAI. https://www.microsoft.com/en-us/research/publication/predicting-depression-via-social-media/

[65] Munmun De Choudhury, Choudhury Michael, and Gamon Scott Counts. [n.d.]. *Happy, Nervous or Surprised? Classification of Human Affective States in Social Media.*

[66] Jose van Dijck. 2014. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance & Society* 12, 2 (May 2014), 197–208. https://doi.org/10.24908/ss.v12i2.4776

[67] J. Doyle, N. Caprani, and R. Bond. 2015. Older adults' attitudes to self-management of health and wellness through smart home data. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. 129–136. https://doi.org/10.4108/icst.pervasivehealth.2015.259279

[68] J. A. Easterbrook. 1959. The effect of emotion on cue utilization and the organization of behavior. *Psychological Review* 66, 3 (May 1959), 183–201. https://doi.org/10.1037/h0047707 Publisher: American Psychological Association.

[69] Laura Eggertson. 2015. Social media embraces suicide prevention. *CMAJ* 187, 11 (Aug. 2015), E333–E333. https://doi.org/10.1503/cmaj.109-5104 Publisher: CMAJ Section: News.

[70] Paul. Ekman. 2003. *Emotions revealed : recognizing faces and feelings to improve communication and emotional life* (1st ed. ed.). Number xvii, 267 p. :. Times Books, New York :. Publication Title: Emotions revealed : recognizing faces and feelings to improve communication and emotional life.

[71] Paul Ekman. 2009. Darwin's contributions to our understanding of emotional expressions. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (Dec. 2009), 3449–3451. https://doi.org/10.1098/rstb.2009.0189

[72] Paul. Ekman and Wallace V. Friesen. 2003. *Unmasking the face : a guide to recognizing emotions from facial clues.* Number xii, 212 p. :. Malor Books, Cambridge, MA :. Publication Title: Unmasking the face : a guide to recognizing emotions from facial clues.

[73] Nicole B. Ellison, Charles Steinfield, and Cliff Lampe. 2007. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication* 12, 4 (2007), 1143–1168. https://doi.org/10.1111/j.1083-6101.2007.00367.x _eprint: https://academic.oup.com/jcmc/article-pdf/12/4/1143/22316419/jjcmcom1143.pdf.

[74] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I always assumed that I wasn't really that close to [her]": Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 153–162. https://doi.org/10.1145/2702123.2702556

[75] Gunther Eysenbach. 2009. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. *Journal of Medical Internet Research* 11, 1 (2009), e11. https://doi.org/10.2196/jmir.1157 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[76] Jessica L. Feuston and Anne Marie Piper. 2018. Beyond the Coded Gaze: Analyzing Expression of Mental Health and Illness on Instagram. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 51 (Nov. 2018), 21 pages. https://doi.org/10.1145/3274320

[77] Jessica L. Feuston and Anne Marie Piper. 2018. Beyond the Coded Gaze: Analyzing Expression of Mental Health and Illness on Instagram. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–21. https://doi.org/10.1145/3274320

[78] Jessica L. Feuston and Anne Marie Piper. 2019. Everyday Experiences: Small Stories and Mental Illness on Instagram. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300495

[79] Jessica L. Feuston and Anne Marie Piper. 2019. Everyday experiences: Small stories and mental illness on Instagram. In *CHI 2019 - Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing

Machinery. https://doi.org/10.1145/3290605.3300495
[80] Jessica L. Feuston, Alex S. Taylor, and Anne Marie Piper. 2020. Conformity of Eating Disorders through Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–28. https://doi.org/10.1145/3392845
[81] Casey Fiesler, Cliff Lampe, and Amy S. Bruckman. 2016. Reality and Perception of Copyright Terms of Service for Online Content Creation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1450–1461. https://doi.org/10.1145/2818048.2819931
[82] Casey Fiesler and Nicholas Proferes. [n.d.]. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4, 1 ([n. d.]). https://doi.org/10.1177/2056305118763366
[83] Janet Finch. 1987. THE VIGNETTE TECHNIQUE IN SURVEY RESEARCH. *Sociology* 21, 1 (1987), 105–114. http://www.jstor.org/stable/42854387 Publisher: Sage Publications, Ltd.
[84] FINDER. 2020. Emotion recognition technology in the financial sector – Curse or blessing? https://thefinderproject.eu/2020/01/07/emotion-recognition-technology-in-the-financial-sector-curse-or-blessing/
[85] Elizabeth Ford, Keegan Curlewis, Akkapon Wongkoblap, and Vasa Curcin. 2019. Public Opinions on Using Social Media Content to Identify Users With Depression and Target Mental Health Care Advertising: Mixed Methods Survey. *JMIR Mental Health* 6, 11 (2019), e12942. https://doi.org/10.2196/12942 Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada.
[86] Brett M. Frischmann and Evan Selinger. 2018. *Re-engineering humanity*. Cambridge University Press, Cambridge, United Kingdom ; New York, NY.
[87] Tarleton Gillespie. 2014. The Relevance of Algorithms. In *Media Technologies*, Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot (Eds.). The MIT Press, 167–194. https://doi.org/10.7551/mitpress/9780262525374.003.0009
[88] Susan E Gindin. 2009. Nobody reads your privacy policy or online contract: Lessons learned and questions raised by the FTC's action against Sears. *Nw. J. Tech. & Intell. Prop.* 8 (2009), 1.
[89] Tasha Glenn and Scott Monteith. 2014. Privacy in the Digital World: Medical and Health Data Outside of HIPAA Protections. *Current Psychiatry Reports* 16, 11 (Sept. 2014), 494. https://doi.org/10.1007/s11920-014-0494-4
[90] Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. 2018. Ethics and Artificial Intelligence: Suicide Prevention on Facebook. *Philosophy & technology* 31, 4 (2018), 669–684. https://doi.org/10.1007/s13347-018-0336-0 Place: Dordrecht Publisher: Springer Science and Business Media LLC.
[91] John F. Gunn III and David Lester. 2013. Using google searches on the internet to monitor suicidal behavior. *Journal of Affective Disorders* 148, 2 (June 2013), 411–412. https://doi.org/10.1016/j.jad.2012.11.004
[92] Mario Haim, Florian Arendt, and Sebastian Scherr. 2017. Abyss or Shelter? On the Relevance of Web Search Engines' Search Results When People Google for Suicide. *Health Communication* 32, 2 (Feb. 2017), 253–258. https://doi.org/10.1080/10410236.2015.1113484 Publisher: Routledge _eprint: https://doi.org/10.1080/10410236.2015.1113484.
[93] Oliver L. Haimson, Jed R. Brubaker, Lynn Dombrowski, and Gillian R. Hayes. 2015. Disclosure, Stress, and Support During Gender Transition on Facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1176–1190. https://doi.org/10.1145/2675133.2675152
[94] Blake Hallinan, Jed R Brubaker, and Casey Fiesler. 2020. Unexpected expectations: Public reaction to the Facebook emotional contagion study. *New Media & Society* 22, 6 (June 2020), 1076–1094. https://doi.org/10.1177/1461444819876944 Publisher: SAGE Publications.
[95] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173582
[96] Alex Hanna and Meredith Whittaker. [n.d.]. Opinion: Timnit Gebru's Exit From Google Exposes a Crisis in AI. *Wired* ([n. d.]). https://www.wired.com/story/timnit-gebru-exit-google-exposes-crisis-in-ai/
[97] Carl L. Hanson, Scott H. Burton, Christophe Giraud-Carrier, Josh H. West, Michael D. Barnes, and Bret Hansen. 2013. Tweaking and Tweeting: Exploring Twitter for Nonmedical Use of a Psychostimulant Drug (Adderall) Among College Students. *Journal of Medical Internet Research* 15, 4 (2013), e62. https://doi.org/10.2196/jmir.2503 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.
[98] Aya Hassouneh, A.M. Mutawa, and Murugappan M. 2020. Development of a Real-Time Emotion Recognition System Using Facial Expressions and EEG based on machine learning and deep neural network methods. *Informatics in Medicine Unlocked* 20 (2020), 100372. https://doi.org/10.1016/j.imu.2020.100372

[99] Andrew C. High, Anne Oeldorf-Hirsch, and Saraswathi Bellur. 2014. Misery rarely gets company: The influence of emotional bandwidth on supportive communication on Facebook. *Computers in Human Behavior* 34 (May 2014), 79–88. https://doi.org/10.1016/j.chb.2014.01.037

[100] Chris Jay Hoofnagle and Jan Whittington. 2013. Free: accounting for the costs of the internet's most popular price. *UCLA L. Rev.* 61 (2013), 606.

[101] Kimberly A Houser and W Gregory Voss. 2018. GDPR: The end of Google and facebook or a new paradigm in data privacy. *Rich. JL & Tech.* 25 (2018), 1.

[102] Elise Hu. [n.d.]. Facebook Manipulates Our Moods For Science And Commerce: A Roundup. https://www.npr.org/sections/alltechconsidered/2014/06/30/326929138/facebook-manipulates-our-moods-for-science-and-commerce-a-roundup

[103] Kit Huckvale, Svetha Venkatesh, and Helen Christensen. 2019. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. *npj Digital Medicine* 2, 1 (Sept. 2019), 1–11. https://doi.org/10.1038/s41746-019-0166-1 Number: 1 Publisher: Nature Publishing Group.

[104] Rhidian Hughes. 1998. Considering the vignette technique and its application to a study of drug injecting and HIV risk and safer behaviour. *Sociology of Health & Illness* 20, 3 (1998), 381–400.

[105] Sarah E Igo. 2020. *The known citizen: a history of privacy in modern America.* OCLC: 1111377193.

[106] Digital Life Initiative. 2018. Facebook and Google Are the New Data Brokers. https://www.dli.tech.cornell.edu/post/facebook-and-google-are-the-new-data-brokers

[107] A. Jaimes, D. Gatica-Perez, N. Sebe, and T. S. Huang. 2007. Guest Editors' Introduction: Human-Centered Computing–Toward a Human Revolution. *Computer* 40, 5 (May 2007), 30–34. https://doi.org/10.1109/MC.2007.169

[108] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. AI-Mediated Communication: How the Perception That Profile Text Was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300469

[109] Dr William James. 2013. *What is an Emotion?* Simon and Schuster.

[110] Ayn de Jesus. [n.d.]. Artificial Intelligence in Video Marketing - Emotion Recognition, Video Generation, and More. https://emerj.com/ai-sector-overviews/artificial-intelligence-for-video-marketing-emotion-recognition-video-generation-and-more/

[111] Bobbie Johnson and Las Vegas. 2010. Privacy no longer a social norm, says Facebook founder. *The Guardian* (Jan. 2010). https://www.theguardian.com/technology/2010/jan/11/facebook-privacy

[112] Kimberly Barsamian Kahn, Phillip Atiba Goff, J. Katherine Lee, and Diane Motamed. 2016. Protecting Whiteness: White Phenotypic Racial Stereotypicality Reduces Police Use of Force. *Social Psychological and Personality Science* 7, 5 (July 2016), 403–411. https://doi.org/10.1177/1948550616633505 Publisher: SAGE Publications Inc.

[113] Olivia J. Kirtley and Rory C. O'Connor. 2020. Suicide prevention is everyone's business: Challenges and opportunities for Google. *Social Science & Medicine* (Jan. 2020), 112691. https://doi.org/10.1016/j.socscimed.2019.112691

[114] Funda Kivran-Swaine, Jeremy Ting, Jed Richards Brubaker, Rannie Teodoro, and Mor Naaman. [n.d.]. Understanding Loneliness in Social Awareness Streams: Expressions and Responses. ([n. d.]), 10.

[115] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111, 24 (June 2014), 8788. https://doi.org/10.1073/pnas.1320040111

[116] Elizabeth M. Lawrence. 2017. Why Do College Graduates Behave More Healthfully than Those Who Are Less Educated? *Journal of Health and Social Behavior* 58, 3 (2017), 291–306. https://doi.org/10.1177/0022146517715671

[117] Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. Emotion and decision making. *Annual Review of Psychology* 66 (Jan. 2015), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

[118] Kevin Litman-Navarro. 2019. Opinion | We Read 150 Privacy Policies. They Were an Incomprehensible Disaster. *The New York Times* (June 2019). https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html,https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html

[119] Jorge Lopez-Castroman, Bilel Moulahi, Jérôme Azé, Sandra Bringay, Julie Deninotti, Sebastien Guillaume, and Enrique Baca-Garcia. 2020. Mining social networks to improve suicide prevention: A scoping review. *Journal of Neuroscience Research* 98, 4 (2020), 616–625. https://doi.org/10.1002/jnr.24404 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jnr.24404.

[120] Brian Lubars and Chenhao Tan. 2019. Ask Not What AI Can Do, But What AI Should Do: Towards a Framework of Task Delegability. *CoRR* abs/1902.03245 (2019). arXiv:1902.03245 http://arxiv.org/abs/1902.03245

[121] Jennifer Lynch. 2018. Face Off: Law Enforcement Use of Face Recognition Technology. https://www.eff.org/wp/law-enforcement-use-face-recognition

[122] Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. 2010. Social sensing for epidemiological behavior change. In *Proceedings of the 12th ACM international conference on Ubiquitous computing (UbiComp '10)*. Association

for Computing Machinery, New York, NY, USA, 291–300. https://doi.org/10.1145/1864349.1864394

[123] M. Magdin, T. Sulka, J. Tomanova, and M. Vozar. 2019. Voice Analysis Using PRAAT Software and Classification of User Emotional State. *International Journal of Interactive Multimedia and Artificial Intelligence* 5, 6 (Sept. 2019), 33–. http://link.gale.com/apps/doc/A600161373/AONE?u=umuser&sid=zotero&xid=104652f4

[124] Lydia Manikonda and Munmun De Choudhury. 2017. Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 170–181. https://doi.org/10.1145/3025453.3025932

[125] Mason Marks. 2019. Artificial Intelligence-Based Suicide Prediction. *ARTIFICIAL INTELLIGENCE* (2019), 24.

[126] Kirsten E. Martin and Helen Nissenbaum. 2015. *Measuring Privacy: An Empirical Test Using Context To Expose Confounding Variables*. SSRN Scholarly Paper ID 2709584. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.2709584

[127] Eli Zimmerman Twitter Eli has been eagerly pursuing a journalistic career since he left the University of Maryl, 's Philip Merrill School of Journalism Previously, Eli was a staff reporter for medical trade publication Frontline Medical News, Where He Experienced the Impact of Continuous Education, evolving teaching methods through the medical lens When not in the office, and Eli is busy scanning the web for the latest podcasts or stepping into the boxing ring for a few rounds. [n.d.]. GoGuardian Develops a New AI-Enabled Cloud Filter for K–12 Schools. https://edtechmagazine.com/k12/article/2019/02/goguardian-develops-new-ai-enabled-cloud-filter-k-12-schools

[128] John McCormick. 2019. What AI Can Tell From Listening to You. *Wall Street Journal* (April 2019). https://www.wsj.com/articles/what-ai-can-tell-from-listening-to-you-11554169408

[129] Kwame McKenzie and Kamaldeep Bhui. 2007. Institutional racism in mental health care. *BMJ : British Medical Journal* 334, 7595 (March 2007), 649–650. https://doi.org/10.1136/bmj.39163.395972.80

[130] Andrew McStay. [n.d.]. Emotional AI: The Rise of Empathic Media. https://www.researchgate.net/publication/326294717_Emotional_AI_The_Rise_of_Empathic_Media

[131] Karen McVeigh. 2014. Samaritans Twitter app identifying user's moods criticised as invasive. *The Guardian* (Nov. 2014). https://www.theguardian.com/society/2014/nov/04/samaritans-twitter-app-mental-health-depression

[132] Robinson Meyer. 2014. Everything We Know About Facebook's Secret Mood Manipulation Experiment. https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/ Section: Technology.

[133] Jude Mikal, Samantha Hurst, and Mike Conway. 2016. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. *BMC medical ethics* 17, 1 (2016), 22.

[134] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the Language of Schizophrenia in Social Media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Denver, Colorado, 11–20. https://doi.org/10.3115/v1/W15-1202

[135] Dan Muriello, Lizzy Donahue, Danny Ben-David, Umut Ozertem, and Reshef Shilon. 2018. Under the hood: Suicide prevention tools powered by AI. https://engineering.fb.com/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/ Section: ML Applications.

[136] Mark Myslín, Shu-Hong Zhu, Wendy Chapman, and Mike Conway. 2013. Using Twitter to Examine Smoking Behavior and Perceptions of Emerging Tobacco Products. *Journal of Medical Internet Research* 15, 8 (2013), e174. https://doi.org/10.2196/jmir.2534 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[137] M. Namara, H. Sloan, P. Jaiswal, and B. P. Knijnenburg. 2018. The Potential for User-Tailored Privacy on Facebook. In *2018 IEEE Symposium on Privacy-Aware Computing (PAC)*. 31–42. https://doi.org/10.1109/PAC.2018.00010

[138] Natasha Singer. 2018. Creepy or Not? Your Privacy Concerns Probably Reflect Your Politics. *New York Times (Online)* (April 2018). https://www.nytimes.com/2018/04/30/technology/privacy-concerns-politics.html Journal Abbreviation: New York Times (Online) Publisher: New York Times Company.

[139] James P. Nehf. 2010. The FTC's Proposed Framework for Privacy Protection Online: A Move toward Substantive Controls or Just More Notice and Choice Electronic Commerce Law. *William Mitchell Law Review* 37, 4 (2010), 1727–1744. https://heinonline.org/HOL/P?h=hein.journals/wmitch37&i=1737

[140] Casey Newton. 2020. How Facebook is preparing for a surge in depressed and anxious users. https://www.theverge.com/2020/3/19/21185204/facebook-coronavirus-depression-anxiety-content-moderation-mark-zuckerberg-interview

[141] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington Law Review* 79, 1 (Feb. 2004), 119–157.

[142] Frank A. Pasquale. 2020. More Than a Feeling. https://reallifemag.com/more-than-a-feeling/

[143] Jessica A Pater, Oliver L Haimson, Nazanin Andalibi, and Elizabeth D Mynatt. 2016. "Hunger Hurts but Starving Works" Characterizing the Presentation of Eating Disorders Online. In *Proceedings of the 19th ACM Conference on*

Computer-Supported Cooperative Work & Social Computing. 1185–1200.

[144] Janice Penni. 2017. The future of online social networks (OSN): A measurement analysis using social media tools and application. *Telematics and Informatics* 34, 5 (Aug. 2017), 498–517. https://doi.org/10.1016/j.tele.2016.10.009

[145] R W Picard, S Papert, W Bender, B Blumberg, C Breazeal, D Cavallo, T Machover, M Resnick, D Roy, and C Strohecker. 2004. Affective Learning — A Manifesto. *BT Technology Journal* 22, 4 (Oct. 2004), 253–269. https://doi.org/10.1023/B:BTTJ.0000047603.37042.33

[146] Harold A. Pollack and Keith Humphreys. 2020. Reducing Violent Incidents between Police Officers and People with Psychiatric or Substance Use Disorders. *The ANNALS of the American Academy of Political and Social Science* 687, 1 (Jan. 2020), 166–184. https://doi.org/10.1177/0002716219897057 Publisher: SAGE Publications Inc.

[147] Juan Carlos Quiroz, Elena Geangu, and Min Hooi Yong. 2018. Emotion Recognition Using Smart Watch Sensor Data: Mixed-Design Study. *JMIR Mental Health* 5, 3 (Aug. 2018), e10153. https://doi.org/10.2196/10153

[148] Andrew G. Reece and Christopher M. Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6, 1 (Dec. 2017), 1–12. https://doi.org/10.1140/epjds/s13688-017-0110-z Number: 1 Publisher: SpringerOpen.

[149] Farig Sadeque, Dongfang Xu, and Steven Bethard. 2018. Measuring the Latency of Depression Detection in Social Media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 495–503. https://doi.org/10.1145/3159652.3159725

[150] Najmeh Samadiani, Guangyan Huang, Borui Cai, Wei Luo, Chi-Hung Chi, Yong Xiang, and Jing He. 2019. A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data. *Sensors* 19, 8 (Jan. 2019), 1863. https://doi.org/10.3390/s19081863 Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[151] Andrea Scarantino. 2015. Basic emotions, psychological construction, and the problem of variability. In *The psychological construction of emotion*. Guilford Press, New York, NY, US, 334–376.

[152] Maya Schenwar, Macaré Joe, and Alana Yu-lan Price. 2016. *Who do you serve, who do you protect?: police violence and resistance in the United States*. Haymarket Books, Chicago, Illinois. OCLC: 975049067.

[153] Klaus R. Scherer, Angela. Schorr, and Tom. Johnstone. 2001. *Appraisal processes in emotion : theory, methods, research*. Number xiv, 478 p. : in Series in affective science. Oxford University Press, Oxford ; New York :. Publication Title: Appraisal processes in emotion : theory, methods, research.

[154] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. ACM Press, Atlanta, GA, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[155] Evan Selinger and Woodrow Hartzog. 2016. Facebook's emotional contagion study and the ethical problem of co-opted identity in mediated environments where users lack control. *Research Ethics* 12, 1 (Jan. 2016), 35–43. https://doi.org/10.1177/1747016115579531 Publisher: SAGE Publications Ltd.

[156] Manoj Kumar Sharma, Nisha John, and Maya Sahu. 2020. Influence of social media on mental health: a systematic review. *Current Opinion in Psychiatry* Publish Ahead of Print (July 2020). https://doi.org/10.1097/YCO.0000000000000631

[157] Natasha Singer. 2018. In Screening for Suicide Risk, Facebook Takes On Tricky Public Health Role. *The New York Times* (Dec. 2018). https://www.nytimes.com/2018/12/31/technology/facebook-suicide-screening-algorithm.html

[158] Vivek K. Singh and Rishav R. Agarwal. 2016. Cooperative phoneotypes: exploring phone-based behavioral markers of cooperation. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 646–657. https://doi.org/10.1145/2971648.2971755

[159] Luke Stark and Jesse Hoey. 2020. *The Ethics of Emotion in AI Systems*. Technical Report. OSF Preprints. https://doi.org/10.31219/osf.io/9ad4u

[160] George Szmukler. 2015. Compulsion and "coercion" in mental health care. *World Psychiatry* 14, 3 (Oct. 2015), 259–261. https://doi.org/10.1002/wps.20264

[161] Ziying Tan, Xingyun Liu, Xiaoqian Liu, Qijin Cheng, and Tingshao Zhu. 2017. Designing Microblog Direct Messages to Engage Social Media Users With Suicide Ideation: Interview and Survey Study on Weibo. *Journal of Medical Internet Research* 19, 12 (Dec. 2017), e381. https://doi.org/10.2196/jmir.8729

[162] Marilyn D. Thomas, Nicholas P. Jewell, and Amani M. Allen. 2021. Black and unarmed: statistical interaction between age, perceived mental illness, and geographic region among males fatally shot by police using case-only design. *Annals of Epidemiology* 53 (Jan. 2021), 42–49.e3. https://doi.org/10.1016/j.annepidem.2020.08.014

[163] John Torous, Mathew V Kiang, Jeanette Lorme, and Jukka-Pekka Onnela. 2016. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Mental Health* 3, 2 (May 2016). https://doi.org/10.2196/mental.5165

[164] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing Depression from Twitter Activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3187–3196. https://doi.org/10.1145/

2702123.2702280

[165] Donghyeon Won, Zachary C. Steinert-Threlkeld, and Jungseock Joo. 2017. Protest Activity Detection and Perceived Violence Estimation from Social Media Images. In *Proceedings of the 25th ACM International Conference on Multimedia* (Mountain View, California, USA) *(MM '17)*. Association for Computing Machinery, New York, NY, USA, 786–794. https://doi.org/10.1145/3123266.3123282

[166] Richmond Y. Wong, Deirdre K. Mulligan, and John Chuang. 2017. Using science fiction texts to surface user reflections on privacy. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers (UbiComp '17)*. Association for Computing Machinery, New York, NY, USA, 213–216. https://doi.org/10.1145/3123024.3123080

[167] Richmond Y. Wong, Ellen Van Wyk, and James Pierce. 2017. Real-Fictional Entanglements: Using Science Fiction and Design Fiction to Interrogate Sensing Technologies. (June 2017). https://doi.org/10.1145/3064663.3064682

[168] Hao Yan, Ellen E. Fitzsimmons-Craft, Micah Goodman, Melissa Krauss, Sanmay Das, and Patricia Cavazos-Rehg. 2019. Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention. *The International Journal of Eating Disorders* 52, 10 (2019), 1150–1156. https://doi.org/10.1002/eat.23148

[169] Shoshana Zuboff. 2020. *The age of surveillance capitalism: the fight for a human future at the new frontier of power* (first trade paperback edition ed.). PublicAffairs, New York.

Received October 2020; revised April 2021; accepted May 2021