A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media

Stevie Chancellor
Georgia Tech
Atlanta, GA, US
schancellor3@gatech.edu

Michael L Birnbaum
Northwell Health
Glen Oaks, NY, US
mbirnbaum@northwell.edu

Eric D. Caine
University of Rochester
Rochester, NY, US
Eric_Caine@urmc.rochester.edu

Vincent M. B. Silenzio
University of Rochester
Rochester, NY, US
vincent.silenzio@rochester.edu

Munmun De Choudhury
Georgia Tech
Atlanta, GA, US
munmund@gatech.edu

ABSTRACT

Powered by machine learning techniques, social media provides an unobtrusive lens into individual behaviors, emotions, and psychological states. Recent research has successfully employed social media data to predict mental health states of individuals, ranging from the presence and severity of mental disorders like depression to the risk of suicide. These algorithmic inferences hold great potential in supporting early detection and treatment of mental disorders and in the design of interventions. At the same time, the outcomes of this research can pose great risks to individuals, such as issues of incorrect, opaque algorithmic predictions, involvement of bad or unaccountable actors, and potential biases from intentional or inadvertent misuse of insights. Amplifying these tensions, there are also divergent and sometimes inconsistent methodological gaps and under-explored ethics and privacy dimensions. This paper presents a taxonomy of these concerns and ethical challenges, drawing from existing literature, and poses questions to be resolved as this research gains traction. We identify three areas of tension: ethics committees and the gap of social media research; questions of validity, data, and machine learning; and implications of this research for key stakeholders. We conclude with calls to action to begin resolving these interdisciplinary dilemmas.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; **Social media**; • **Applied computing** → *Psychology*;

KEYWORDS

mental health; ethics; machine learning; algorithms; social media

ACM Reference Format:
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 FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287587

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
FAT '19, January 29–31, 2019, Atlanta, GA, USA*
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6125-5/19/01…$15.00
https://doi.org/10.1145/3287560.3287587

1 INTRODUCTION

Last year, Facebook unveiled automated tools to identify individuals contemplating suicide or self-injury [75, 62]. The company claims that they "use pattern recognition technology to help identify posts and live streams as likely to be expressing thoughts of suicide," which then can deploy resources to assist the person in crisis [75]. Reactions to Facebook's suicide prevention artificial intelligence (AI) are mixed, with some concerned about the use of AI to detect suicidal ideation as well as potential privacy violations [86]. Other suicide prevention AIs, however, have been met with stronger public backlash. Samaritan's Radar, an app that scanned a person's friends for concerning Twitter posts, was pulled from production, citing concerns for data collection without user permission [54], as well as enabling harassers to intervene when someone was vulnerable [4].

Since 2013, a new area of research has incorporated techniques from machine learning, natural language processing, and clinical psychology to categorize individuals' moods and expressed well-being from social media data. These algorithms are powerful enough to infer with high accuracy whether an individual might be suffering from disorders such as major depression [28, 19, 84, 73, 78], postpartum depression [26, 27], post-traumatic stress [21], schizophrenia [60, 6], and suicidality [15, 22]. These algorithms can also reveal symptomatology linked to psychiatric challenges, such as self-harm [89], severity of distress [13], or cognitive distortions [82]. Together, we use the term predicting *mental health status* to describe these mental disorders and related symptomatology.

Computer Science (CS) researchers and clinicians are now poised to learn more about the earliest manifestations of psychiatric disorders through social media data. New insights could prevent the development of latent conditions, mitigate the impact of emerging disorders, or as exemplified by Facebook's new suicide AI, new opportunities to intervene with life-saving assistance. With the rising prevalence of mental disorders [67], many researchers see the benefits of better screening, identification, and intervention assisting to promote better health and well-being worldwide.

However, the examples of suicide prevention AIs demonstrate major concerns for algorithmic development and their implications. This includes new concerns about consent into monitoring or intervention systems and privacy and data management questions. Ethics boards do not have standards for managing social

media research, and the prediction of mental health status raises new questions about consent, vulnerable populations, and online communities. There are also methodological concerns of data collection and bias, validity of these results for clinical assessment, and the application of machine learning methods to predicting mental health status. Furthermore, the lack of consistency with methods across this research space makes this problem more troubling. For implications, actors with many motivations can misuse data and predictions, and amplify the harms of algorithms in reproducing unfair stereotypes and discrimination of individuals with mental disorders. Caused in part by the interdisciplinary intersection of data science, machine learning, psychology, and human-centered computing, unanswered questions emerge around the role of the individual in predictions and managing implications of this research.

As these technologies are developed to detect mental health status, these concerns will grow unless we rectify these problems. We stand to gain much from this research – in better understanding and making interventions in mental health. Addressing these concerns will resolve questions around rigorous science in the area, benefit clinical research, and safeguard well-being for individuals and society. Many of these concerns are not limited to just mental health and social media and apply to other application domains of these technologies that touch on sensitive issues. In answering these questions, we offer insight into questions on how to ethically and rigorously apply machine learning and AI to sensitive domains such as mental health, and we provide this analysis as a case study for ethics in applied and fair AI.

This work presents a first taxonomy of issues in algorithmic prediction of mental health status on social media data. First, we discuss the gap between ethics committees and participants in such research, on what can be sensitive and sometimes stigmatizing data. Second, we identify tensions in methods and analysis, such as construct validity and bias, interpretability of algorithmic output, and privacy. Finally, we examine implications of this research in benefiting mental health research, challenges faced by key stakeholders, and the risks of designing interventions.

We contextualize these three areas by drawing from prior work in this domain, ethics research around these technological advances, and our experiences conducting this research. In our analysis of each of our three areas, we look to prior work and standards across fields: machine learning (ML), natural language processing (NLP), human computer interaction (HCI), clinical psychiatry, and data science for guidance. We conclude with calls for interdisciplinary action to resolve these dilemmas.

2 STATE OF THE ART IN THE FIELD

The origins of predictive work come from either population-level analyses or studies of generalized and subjective well-being and affect assessment. Borrowing from advances in natural language processing [11] and psychology [71] to represent text as cues of well-being, these studies described mood shifts around political events [7], geographic differences in expressed well-being [79], and the seasonality and temporality of mood variation [38]. In addition to studying generalized well-being, researchers also assessed population happiness both on Twitter [31] and Facebook [51]. Besides establishing that psychological and health states can be inferred from this data, these findings show that people use social media to discuss their personal mood and activities honestly and candidly instead of their idealized versions [3]. Complementary to this research were studies in public health measurement with online data, termed "infodemiology." [33] This famously includes the use of human-generated data to predict influenza outbreaks through search engines [36]. Researchers also used social media data to track the spread of disease [76] and to analyze other ailments on population-scale user bases from Twitter [70].

Soon after these studies were the first predictive works on the mental health states of individuals, beginning with depression. In 2013, De Choudhury et al. used clinically validated depression measures to find Twitter users who tested for major depressive disorder [28]. They developed a model that could predict if someone had depression with 70% accuracy. Around the same time, Park et al. developed a mixed methods approach to understand how Facebook use corresponded to clinical scales for depression [68]. In 2014, Coppersmith et al. used self-reported disclosures of depression diagnosis on Twitter ("I was diagnosed with depression on...") to classify individuals suffering with depression, contrasting their language with those who do not self-report such diagnoses [19]. De Choudhury et al. also sought to identify new mothers who might be suffering from postpartum depression using Facebook and Twitter data [26, 27]. After these works, researchers began to replicate, extend, generalize, and improve on these findings [63, 78, 74] and in different cultural contexts and social media sites, beyond just English-speaking Twitter [84, 88].

From these seminal works on depression, new studies have investigated new psychiatric disorders, new social network platforms, and new modalities. Research has examined other disorders, such as post-traumatic stress disorder [21], anxiety [81], schizophrenia [6, 60], eating disorders [13, 14, 25], and suicidal ideation [30, 29, 39]. Work also now explores the symptomatology of mental disorders, such as the severity of mental illness [13, 78] and stress connected to mental health [55]. Datasets too have expanded to social networks other than Twitter and Facebook, like Sina Weibo [46], Instagram [73, 13], Tumblr [14, 82], and Reddit [77, 37]. Modalities other than text are now analyzed for their signals of mental health status. Automated image analysis can identify self-harm photos on Flickr [89], signs of depression through Instagram images [73], and mental health disclosures on Reddit [56]. Finally, new data sources have begun to supplement social media data, like active and passive sensing technologies [77].

Ethical Considerations in Existing Research. Overall, the field of deriving algorithmic predictions of individuals' mental health status is a growing area of research interest across sub-disciplines of CS and is gaining traction in relevant domains [6, 15]. Most, though not all, of this works touches on ethical and methods challenges as well as steps researchers take to mitigate risks to individuals whose data is analyzed. Many papers include explicit notes about obfuscating sensitive and personally identifiable information [73, 77], data de-identification [22], involvement of domain experts for responsible data handling and curation [13], the need for ethical and privacy sensitivity in technology powered by algorithmic inferences [68, 25], quality of inferences among potential stakeholders [26], and the need for cross-disciplinary collaboration and dialogue to prevent misuse and misinterpretation of algorithmic outcomes [14].

Notable Gaps. However, there are no accepted guidelines to navigate these challenges; decisions by a particular research team omitted from papers are often invisible to the community, leading to difficulties in normalizing ethical considerations. Given the vulnerability and sensitivity of the population and the topic, we find this concerning. Discussions of consent, validity, underlying bias from data collection techniques, or machine learning model selection is very limited, even though applying algorithms in practical scenarios features prominently as an end goal of this research. To frame a new set of interdisciplinary ethical guidelines in this emergent research area, we look to these works to inform our analysis.

3 INSIGHTS FROM ETHICS RESEARCH

Complementary to this work is a long history investigating the ethics of computing technology on broader domains. In fact, some of the gaps we note above, such as participant consent, role of ethics boards, and challenges to autonomy and privacy, have been discussed at length in these works. Given the growing significance of machine learning and algorithms in different domains, this field has received renewed attention both within the FAT* community [50, 24] as well as the field of "critical algorithms" [8, 58, 35]. We provide a brief overview of relevant research in three spaces: social media research ethics, public health research, and critical data studies.

Social Media Research Ethics. Ample research has addressed issues in social media and ethics, as early as 2004 [47]. Moving into the age of "big data," scholars are considering how new methods and data aggregation techniques impact individuals involved in this research. Hargittai analyzed the snowballing effects that of unintended biased sampling of social media data on big data analyses [40]. Zimmer has examined ethical use of Facebook data [93] and proposed a topology of ethical issues from Twitter research [94]. Finally, Olteanu et al. considered the methodological challenges of mining social media for information, including issues of internal and external validity, data curation, and methods [66].

Public Health and Ethics. Second, we look to the history of public health research, social media, and ethics for population-scale predictions of disease and disorders. Dredze and Paul consider social media research for public health, focusing on end-to-end consideration of study design, identifying target conditions, methods, and ethics [69]. Next, Conway and Connor address advances and ethics of population-scale predictions of mental health, providing an overview of the field and reflecting on how "big data" methods like machine learning and NLP facilitate surveillance of mental health for populations [18]. Metaphors for social surveillance of public health have been proposed, like Vayena et al.'s "digital epidemiology" to understand ethical obligations of researchers using public data [85]. Horvitz and Mulligan analyzed the potential legal, privacy, and data protection issues of big data analysis for well-being [45]. Norval and Henderson unpack various theories of privacy to analyze whether informed consent should be gathered in social media health research for patient information [65], while Mikal et al. used focus groups to understand users perceptions of social media data use for mental health research [59].

In NLP, Benton et al. recently considered the protocols for ethical social media health research from their own experiences in the field [5]. Their work discusses the ethical contention surrounding the use of public social media data for population health inference and its exemption from review by U.S. Institutional Review Boards (IRBs). Stylistically, this work is closest to our position, although the ethical guidelines provided by Benton et al. are geared toward public health needs, not individualized predictions

Critical Data Studies. Finally, the intersection of critical technology research and big data has led to "critical data studies," providing useful metaphors and case studies on the impacts of big data research. In an early work, boyd and Crawford push the new field of data science to critically consider its methods [8]. In response to the failure of Google Flu Trends, Lazer et al. cautions researchers to be cautious in applying predictive techniques [53]. Foucault-Welles brings light to the discriminatory impacts of aggregating analysis of social data that erases differences of minority groups [90]. Metcalf and Crawford discuss the difficulties of using other research relationship metaphors (such as the physician-patient metaphor) to illuminate how data researchers could conceptualize their users as more than just data sources [58].

These three perspectives discuss important concerns: participant consent [5, 58] and contextual data integrity [85, 90]; data protection, anonymization, and privacy [17, 45, 94, 93]; methodological rigor [66, 70, 53]; bias and validity [40, 66]; and implications of the research for different stakeholders [8, 18]. Drawing from these two larger domains – the state-of-the-art on mental health status prediction and surrounding discussion – we identify three areas of tension that encapsulate concerns in this research area.

4 THREE AREAS OF ETHICAL TENSION

Among the areas of ethical tension identified above, first, we address the research design and approval stage of the research. We consider what is ethical to study, if the work deserves ethics board approval, and to what extent we treat social media users as research subjects in these studies. Second, we examine methodological concerns, like feature generation and algorithm selection. Finally, we consider the implications of what these predictions might mean for clinicians, researchers, and other key stakeholders in this space.

4.1 Participants and Research Oversight

Reacting to unethical behavior in medical and psychological experiments in the 1940s and 1950s, many countries have adopted ethical research standards for human subjects research. This standards manifest in an ethics committee, whether that be an IRB, Federalwide Assurance-certified ethics board, European Union (EU) ethics committees, and corporate internal review committees. Researchers and clinicians must also follow legal requirements to protect the dignity and privacy of individuals. In the United States, the Belmont Report and accompanying Common Rule legislation set protocols for human subject research which receives federal funding [72]. Further, the Health Insurance Portability and Accountability Act (HIPAA) protects privacy of patients in clinical relationships with doctors in the U.S and privacy rights of medical records [83], with similar protections in other countries [48].

Guided by the principles of respect, beneficence, and justice, ethics research boards, e.g., U.S. IRBs, deliberately transform people into "research subjects" in scientific inquiry; this transformation prescribes people with certain rights, protections, and obligations that must be protected [58]. In clinical studies, this obligation is at the forefront of experimental design [32].

Is predicting mental health status on social media human subjects research? How do we assess harm of this mental health research without the oversight of an ethics committee? In this section, we discuss challenges of predicting mental health status outside a clinical setting using data-driven algorithm, and impacts to participants.

Key Areas of Tension:
(1) The Unclear Role of Ethics Committees
(2) Consent at Scale
(3) Vulnerable Populations and Risk
(4) Contextual Integrity of Communities

The Unclear Role of Ethics Committees. Analysis of publicly visible social media data is often exempt from research protections provided to subjects through ethics committees. These studies are exempted for two primary reasons: one, in large-scale data analyses, there is no interaction or intervention with subjects because the research is observational; two, the data being used was publicly available when collected. Many ethics boards consider social media to be public space synonymous with gathering publicly available data that might be stored in Census records or courthouses.

We find this interpretation consistent across different countries and in different research environments [21, 84, 26]. Researchers will often cite one or both of these principles in their data collection sections – there exists no relationship between researcher and social media user, nor a doctor-patient relationship that would mandate medical privacy guidelines come into play. Studies that do interact with subjects, through surveys of crowdworkers [73] or individuals recruited through word of mouth, advertisements [73] or through apps [68], tend to declare appropriate ethics board approval.

However, predicting mental health states using public social media data emphasizes whether this research should be exempt from ethics committee oversight. Unlike in public health [5], predicting mental health states, even if with public data, borders on medical diagnosis, such as predicting the presence of schizophrenia. Research is more than just the "sum of its parts", and extensive secondary analysis can be done from traces of social media data [21, 73]. Mental health is a complex and sensitive area that can be isolating and stigmatizing [23], and harm can be difficult to evaluate, especially in second-order impacts. Is this research human subjects research? How should ethics boards handle this new research paradigm?

Consent at Scale. In traditional human subjects research, participant pools rarely exceeds several hundred. This is because inference about mental health states could only be learned through clinician-patient relationships or lab studies that naturally limits the subject pool. By consenting into this research, participants are aware that they are part of research and therefore being surveilled. Consent could meaningfully be gathered from participants, and served as an important signal for participation.

Unlike clinical mental health studies, social media datasets can contain millions of public posts [37], and user accounts regularly exceed the hundreds of thousands [13] – obtaining consent at this scale is pragmatically impossible. However, there are tensions between the infeasibility of obtaining consent and conducting analysis about mood and well-being on social media. This emerged in scrutinized experimental studies of Facebook data [52], where researchers manipulated the mood of millions of Facebook users without consent. In fact, a recent survey study, though not specific to the mental health domain, found that few social media users were aware that their public content could be used by researchers, and the majority felt that researchers should not be able to use tweets without consent [34]. Essentially, passively collecting data transforms its initial purpose, and we miss essential details of individuals' experiences and symptomatology that may be gained from clinical relationships. Is consent necessary in these contexts, and if so, what is meaningful positive or negative consent?

Vulnerable Populations and Risk. Vulnerable populations, such as prisoners, expectant mothers, and minors require additional protocol to protect participants in the U.S. IRB system [41]. Even riskier research topics, such illegal behaviors, are protected with additional scrutiny. For example, the National Institutes of Health releases certificates of confidentiality that prevents research data from release to anyone, including government authorities [41].

No restrictions exist for studies of public social media users, no matter how vulnerable the population may be. For example, the median age of onset for eating disorders is between 18 and 21 [57]. Given that demographic attributes such as age are inferrable from social media language [80], should we research online eating disorder communities, knowing a large subset of these individuals are likely minors [13, 14]? When should data scientists consider vulnerable populations, and how should we protect this data?

Additionally, ethics boards mandate that researchers take actions to protect against risks that a study may cause for mental health. Many clinical studies include a risk management protocol, where participants identified by the research team to be at an elevated mental health risk can be directed to appropriate forms of help and support resources. Researchers can also intervene to stop participation in scientific research if the subject or research team believe the harms outweigh the benefits.

Even in studies without directed interventions, the presence of researchers in communities could be triggering for individuals with mental disorders, e.g. individuals dealing with schizophrenia and fear of mass surveillance may be upset by the knowledge that researchers are tracking their behaviors, even if for beneficial outcomes. Protocols for risk management and drop outs are missing or unimplemented in social media research on mental health. There is no insight into what happens when users "drop out" of social media participation [14], which is a close proxy to withdrawing consent. Are they switching accounts, exiting the platform entirely, or is their mental health state dire? Should we provide information to participants who may be in a dire mental health state?

Contextual Integrity of Communities. Although online communities may post publicly to find support for anxiety [81] to suicidality [30], it is unclear whether social media users understand if their data can be surveilled as they discuss sensitive issues. Behavior in these communities indicate that these groups may have no intention of being discovered by others [13], and they may outright refuse participation in research [47]. When asked directly if users were comfortable with predicting depression with their Twitter profiles, comfort with such research is decisively mixed [59, 34].

Are we violating community norms with these observations? We draw from the notion of "contextual integrity" proposed by Helen Nissenbaum in understanding privacy violations [64], and a related follow-up by Zimmer about contextual gaps in big data

research [92]. Zimmer argues that these gaps cause violations of "normative bounds of how information flows within specific contexts." [92] Is is appropriate to observe online health communities for research if it violates this contextual integrity? What about benign discussions on personal social media accounts?

As Bruckman recommends, one way to resolve this contextual gap is by asking for permission through community leaders [9], which is feasible for Reddit or public Facebook groups. However, most research is done on Twitter data, where no formalized community structure exists, and those that do (like hashtags) are amorphous. Must we ask for consent in these scenarios to maintain contextual integrity, and if so, how would we do this?

4.2 Validity, Interpretability, and Methods

The diversity of fields this research pulls from as well as the venues it publishes in brings many methods questions to the forefront of this work. However, there are documented inconsistencies and unanswered questions in this space (ref. section 2). In this section, we discuss ethical tensions arising from the validity and rigor (or the lack thereof) of new algorithms that infer mental health state.

Key Areas of Tension:

(1) Construct Validity
(2) Data Bias
(3) Algorithmic Interpretability
(4) Performance Tradeoffs
(5) Data Protection and Anonymization

Construct Validity. The American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM) is the best resource for identifying psychiatric symptoms and classifying mental disorders [1]. With over 60 years of empirical support, the DSM guides clinicians and researchers to make accurate psychiatric diagnoses using tested and validated constellations of symptoms and experiences obtained through clinical interviews.

Moreover, clinically and psychometrically validated scales measure the presence and severity of mental disorders, such as the Patient Health Questionnaire (PHQ) or the Generalized Anxiety Disorder scale (GAD-7). It is unclear if mapping these scales to digital contexts validly reproduces results. Further, the complexities of patient-clinician interactions make rote application of DSM guidelines to online social media data unclear, e.g. DSM guideline for diagnostic criteria of certain illness may be misinterpreted, exaggerated, or even lied about on social profiles.

As technology can sense psychiatric symptoms, identify, and potentially diagnose mental illness, we must consider how best to incorporate these tools into clinical practice. How do we map symptom assessment techniques to social media data in a way that preserves its validity? Is it ethical to use mappings of traditional symptomatology or non-traditional ways to predict mental health?

Related to this is valid gold standard labels of mental health status, or "ground truth." For prediction tasks in this space, gathering ground truth data measure the target/predictor variable (mental health status); it is therefore a crucial part of the research process and impacts the quality of the algorithms that are built. There are several standard approaches in the research on assessing ground truth of mental health status, including self-disclosure of mental health state [19, 21, 87], specific hashtag use [37, 14], and community participation [81]. Other styles directly recruit participants and administer screeners, then collect social media data of these participants [28, 73]. Most studies do not include clinical annotation; however, new approaches incorporate clinicians directly in labeling ground truth [6] or validating the accuracy of other sources [13]. These approaches will vary, depending on the research question and study design.

However, there is no guidance on how to select the correct ground truth collection procedure, or whether clinicians are necessary to this process. Are we measuring the phenomenon we argue we are measuring? Are certain kinds of measurement more appropriate for different scenarios? To prevent misinterpretation of the inferences, must we involve clinicians to assess ground truth states?

Data Bias. Bias is a concern for any project; for mental health status prediction, bias is worrisome for the perceived validity and quality of research output. We focus on population biases in datasets (for an excellent analysis of bias, see Olteanu et al.'s survey [66]).

Population bias refers to differences in characteristics between samples in a dataset and those of the target population we intend to measure [66]. The individuals in our datasets (those with a certain mental health status *on social media*) are a subset of the target population (those with a certain mental health status). By gathering data from social media, we bias our data to those who use social media, meaning it is likely a younger and more technologically literate sample than the population as a whole [66].

For mental health status, this bias can manifest in unique ways, leading to ethical lapses and challenges. One well-grounded source of ground truth data is self-reported, diagnosed mental health status (e.g. "I was diagnosed with schizophrenia"). This was pioneered by Coppersmith et al. to unobtrusively identify those with mental disorders [19], and has been validated and used in subsequent projects [21, 60, 6]. By sampling those who publicly self-disclose their mental health diagnoses, this subsample has at least two biases. First, these individuals have (likely) been diagnosed with a mental disorder, meaning they are likely to have sought professional treatment to receive those diagnoses. Second, they are comfortable enough to disclose their mental health status to others, meaning that their forms of sharing could be different from others.

We acknowledge that bias is impossible to avoid in any sampled dataset; however, unaccounted bias can cause latent problems, especially when inferences are incorporated in real life situations. How should we sample and correct for bias? How do we handle these biases in generalizing our results to new mental health statuses, social networks, or contexts?

Algorithmic Interpretability. Next, we discuss ethical challenges arising from a need for algorithmic interpretability and performance [43]. On one end of the spectrum are interpretable models, as in many types of regression models like generalized linear or logistic regressions. As input, these models take intuitive features, derived from social media behavior, known symptomatology [16], or innovations in sub-domains like character n-grams in NLP [20, 22]. As output, these models produce easy-to-understand metrics of model fit and coefficients and probabilities of salient predictors. A strength of these models is that they are easily interpreted by clinicians and stakeholders who may not have technical expertise in algorithmic interpretation, especially when matched to known

symptomatology. However, interpretable models have been known to suffer from poor performance [55, 84, 28]. Regressions and similar algorithms are also limited by data modality, as they do not handle image and video data without extensive preprocessing. Sacrificing performance in the name of interpretability limits applications to applied research. Simply discovering relationships between predictors and outcomes (e.g., risk to a certain mental illness) can be insightful to stakeholders like clinicians; however, it remains unclear how imprecise insights can be actionable during risky situations.

On the other hand, deep learning techniques have emerged as state of the art for powerful and accurate models in prediction tasks. Trained on millions of data points, these algorithms can "effortlessly" outperform other models, handle images and audio, and can intuit features out of the data without human supervision. Performance using deep learning techniques has seen noticeable improvements in predictive power in this space [37, 55]. However, deep learning has a key limitation – they do not produce intelligible feature sets for human understanding [10]. These algorithms are "black boxes," producing impressive results but providing little insight into how the algorithm made its decision. This can make relevant stakeholders in the process, concerned about adopting these algorithms into practical scenarios. Opaque models runs the risk of not only misconstrued and biased conclusions on sensitive data, but also can lead to poor accountability to abide by ethical research principles as well as correcting algorithms when they fail to predict correct outcomes.

These models also challenge human interpretability of their outcomes. How do we handle results that might not align with our clinically-grounded understanding of mental health? These insights might propel research into new areas of signs of mental illness; but, they may also be *red herrings*, providing false hope when in fact the algorithm has latched onto qualities of a particular training set. Multiclass predictions complicate this when they discretize mental health in mutually exclusive binaries (e.g. anxiety or not; depression or not) [19]. The clinical literature overwhelmingly points to mental disorders as frequently co-morbid, and disorders can manifest over a continuous spectrum instead of clearly delineated outputs [1]. Existing algorithmic approaches are often not subtle enough to model this continuum or incorporate interactions between disorders and self-reported symptoms, leading to "artificial" notions of risk.

Performance Tradeoffs. Risks of error in predicting mental health status should be addressed, especially when these algorithms may be used in consumer-facing intervention systems.

False positives, or incorrectly identifying the presence of a mental health status, could cause dramatic consequences for individuals who are the subject of such errors. Many mental disorders are stigmatizing and embarrassing, and being labeled as "disordered" can damage someone's self-esteem, employment prospects, and reputation [23], as was the case of Samaritan's Radar [54]. Depending on implementation, false positives can also cause undue stress on individuals who may now believe something is wrong with them, perhaps stifling their sharing on social platforms in the future. When used in scenarios like content moderation or engagement with a clinician, many false positives may overburden key stakeholders with too many requests to deploy assistance.

On the other hand, a false negative means that mental health status was incorrectly labeled as not having a certain mental health status. Pragmatically, this means no intervention is triggered and no risks for interaction take place. However, in practical use of these systems, false negatives mean that mental health status is missed and may go untreated, as mentioned in prior work [82, 49]. These risks become more concerning when dealing with grave mental health statuses, such as suicidality and psychosis. False negatives also raise responsibility and accountability questions for the results of these algorithms. If being used in functional or practical scenarios, which metric is more important to prioritize? If these algorithms "miss" someone, who is responsible for not intervening? Does this reduce clinician accountability in these scenarios?

Data Sharing and Protection. Even after careful data analysis come risks to privacy for participants. We focus in this section on the risks of data sharing and publication of sensitive information (for excellent overviews of privacy risks, please see Zimmer and Proferes [94], and Horvitz and Mulligan [45]).

Scientists share datasets for reproducibility and consistent benchmarking of new algorithms. However, sharing datasets is complicated by mental health research goals. These datasets are collected under specific circumstances, and users may find issues with context changes. Second, datasets are rarely cleaned for deleted or removed data. In the case of mental health discussions, deleted or removed data could have particularly sensitive data, or data that does not reflect the public perception a person wants to have. How do we manage the joint goals of promoting scientific reproducibility while also protecting participants? What does a benchmarking dataset look like for mental health?

Second is publication of sensitive information such as names, locations, and other personally identifying information. When processing textual social media data, algorithms can occasionally latch onto predictive textual cues; this is amplified when sample size is small. To combat this, researchers have various levels of privacy preservation techniques, such as removing usernames from data before analysis [22] or de-identify algorithmic output later [13]. When should we curate our datasets – pre or post-processing? What are appropriate ways to de-identify data to preserve individual privacy, while maintaining data integrity to promote good science?

A related risk comes from using exemplary social media postings/quotes in papers. Recent work by Ayers et al. found that, of papers that use quotations in papers, over 80% of participants in datasets are able to be reidentified [2]. Other methods, like interview studies, have guidelines on modifying quotations in publications to protect participant identity [9], and we ask similarly: are quotes necessary for demonstrating the validity of the results of the paper? If quotes are needed, what protections can be used for privacy?

4.3 Implications for Stakeholders

Using the perspectives of relevant stakeholders, our final section deals with numerous implications in this research area. We focus on the impacts to researchers in this space, the individuals who are the target of predictions, as well as social networks.

Key areas of tension:

(1) Emotional Vulnerability
(2) Skillset Mismatches
(3) Role of the Clinician
(4) Designing Interventions
(5) Bad Actors and Fairness/Discrimination

Emotional Vulnerability. Researchers and practitioners, especially those from CS, are not often taught how to manage complex emotions when engaging with mental health content. Mental health content can contain graphic and disturbing content, like pictures of self-harm, or detailed discussions of suicide plans [13, 30]. Those who engage with this content can be traumatized by these encounters, and traditional approaches to research design do not take into account the researcher's own emotional well-being [61]. For those who are rarely taught to handle sensitive or emotionally-laden information when annotating and interacting with data, how do we train CS and data scientists to handle the weight of this work?

Skillset Mismatches. There are unique challenges in recognizing and rectifying skill gaps in interdisciplinary research collaborations. Both sets of domain experts must actively work to communicate their research processes and decision-making guidelines this work. As mentioned before (ref. "Algorithmic Interpretability"), algorithmic output can be complex and inscrutable to outsiders. CS researchers are often experts in data collection, feature engineering and model tuning, and performance enhancement. This information needs to be made interpretable to clinicians and other stakeholders with insights into the process. Likewise, CS researchers may lack training in the skills that clinicians traditionally possess. This may be in assessing valid signals of mental health, acquiring ethics board approval, and interpreting signal in datasets.

Some of these decisions may compromise the performance of models, e.g. if a clinician suggests removing a highly predictive feature because it is not clinically relevant to predicting depression, the research team will need to negotiate how to proceed. For these partnerships to blossom, both sets of researchers have to be mindful of making such interpretations accessible to build trust and reliability between collaborators.

Role of the Clinician. Data collected passively/actively or continuously/intermittently may imply different responsibilities for clinicians involved in this research. After entering into a physician-patient relationship, clinicians are bound by the "duty to treat," where they must provide treatment in accordance with their best judgment to their patients. Failing to act on this knowledge would be unethical and potentially illegal. For example, a physician who discovers expressions of suicidal ideation by examining their patient's social media may be bound to treat and therefore intervene.

However, in this field, data is both passive and actively gathered. Information gathered and analyzed passively may not necessarily imply such a strong ethical responsibility for the duty to treat. For example, a clinician annotates an algorithmically-gathered dataset for intent to self-injure and discovers someone states that they plan to committ suicide at a specific date and time – does a clinician have an obligation to intervene? The obligation for intervention here may be weaker, because there is no relationship developed between clinician and social media user.

However, there also exists the "duty to rescue," where a bystander has an obligation to rescue another party in peril. Unlike the duty to treat, the duty to rescue has far more varied interpretations. Does the duty to act or rescue vary depending on the type of professional on the project? In many cases, mental health professionals and computer scientists work in tandem - but what when they work separately? Are computer scientists bound by the duty to rescue if they see someone that says they will harm themselves?

Another question is incorporate these new technologies effectively and ethically into clinical care. How the data is collected, monitored, and presented to the clinical team will alter responsibilities and expectations for clinicians and researchers. For example, research in this space often suggests that insights from this data could be given to clinical care teams [14]. How do we design data interfaces that make sense of these algorithmic predictions for effective insights? How do we not overburden clinicians with large amounts of data and direct their efforts?

Designing Interventions. Another implication is the ability to design interventions, one of the most mentioned applications of this technology in the literature [60, 26, 73]. With suitable performance, the results of these algorithms could provide alerts to help identify moments of crisis, assist in the early identification of mental illness, or avoid risky episodes. The potential for great societal benefit of these prediction algorithms is rooted in these interventions; however, design and implementation of interventions remain a key concern. Outside of clinical interventions, numerous stakeholders are cited as potentially invested in this work, ranging from social networks, crisis hotlines, caregivers, and individual friends to family members. If we detect that a person might be suicidal, should we alert experts or close family members? The automated use of such technologies has been controversial when deployed for Samaritan's Radar [54], but has been better received when driven by human intervention systems on Facebook [75].

There is also risk in alerting individuals of their own mental health status – a piece of information that is inferred algorithmically from passively shared social media data. Are we doing more harm than good by making individuals who are not in a research study aware that they might be suffering from depression or anxiety, thereby alerting them that we have gathered and analyzed their (public) data? These concerns are also connected to issues of managing false positives and false negatives as an important performance tradeoff, as discussed in Section 4.2.

Bad Actors and Fairness/Discrimination. Another issue involves misuse of algorithmic inferences beyond the interests of the individuals themselves by other actors. In one case, the actor has benevolent intentions but misuse the data, or violate the context of what data was gathered. Samaritan's Radar had good intentions of decreasing suicidality, but was poorly received because it enabled other actors to harass or stalk those when they were at their most vulnerable [4]. This can also be seen in automatic screening and text processing systems, like advertising recommendations, which could scan Twitter posts for self-reported diagnoses of mental disorders [19, 21] and send advertisements for prescription drugs. Is this a desirable outcome?

However, researchers have also identified the risks of malintentioned actors using and reproducing the findings in these papers for unsavory purposes [77, 26]. One example could be the use of this research by health insurance agencies to deny coverage for medical care or raise premiums if an individual is detected as "having" postpartum depression yet never sought treatment. Other applications of these algorithms to other prediction systems, like determining credit worthiness for loans or ability to maintain employment status, are possible. In some countries, these predictions are illegal because mental health is a protected class; however, in other cases,

this information is not safeguarded or cleverly designed proxy variables can be engineered to get this information. Can researchers in this space safeguard against bad actors or mitigate these risks?

A related result of these algorithms is discriminatory output – it is possible that the algorithms have a strong sampling bias towards certain groups of people, independent of their mental health status. As mentioned above, social media researchers may be sampling for younger and possibly more affluent audiences by sampling from certain social media data [66]. In their paper about postpartum depression, De Choudhury et al. note that they over-sample Caucasian, affluent women for their data collection and interviews [26], which makes generalizability of this algorithm to other demographics challenging. If we extrapolate our algorithms to these groups, how will we manage unintended biases that might lead to negative and discriminatory repercussions? What impact does this sampling have on predicting on different groups of people, such as those with lower socioeconomic status who do not use social media sites, or older adults with lower rates of social media adoption? Do these algorithms only help the proverbial "rich get richer" by predicting mental health status on groups already likely to seek treatment?

5 CALLS TO ACTION

Research in this area will continue to grow, with new algorithms, data collection means, and new implications for practical use of these algorithms. Even if this taxonomy is not comprehensive, we believe it provides an overview of where to begin to tackle problems, and we are optimistic that the community can work together to solve these challenges. How do we resolve these tensions in predicting mental health status from social media data? Rather than prescribe a set of strict guidelines from our experiences, we call the community to begin to work on these issues. These challenges span both methodological areas in CS as well as topical areas for ethics, privacy, clinical psychiatry, and human-centered design. In this section, we propose three calls to action that could resolve these tensions and inconsistencies in formalized ways.

5.1 Participatory Algorithm Design

> Researchers should include key stakeholders in the research process, including clinicians, social networks, and individuals who are the object of these predictions.

The academic community is already responding to these issues through cross-disciplinary seminars, symposia, and conferences, offering collaborative atmospheres for people to work through these problems. Examples of these venues include the recurrent Computational Linguistics and Clinical Psychology (CLPsych) workshop in NLP; the recurrent Computing and Mental Health symposium at CHI; ML4Health at NIPS in 2017; as well as FAT* itself. These meetings emphasize that interdisciplinary efforts in collegial environments can produce meaningful solutions.

In addition to such partnerships inside the field, CS practitioners should be eager to bring on clinicians and domain experts to this research. Clinical experts provide valuable insights into construct validity, validating and assessing ground truth, correcting for biases, managing risks and privacy tradeoffs, and giving irreplaceable context to algorithmic output. These collaborations are fruitful and have greatly benefit prior research [13, 6, 44]. Other stakeholders, like ethicists, designers, and social media platform owners, should be included as well, as they both offer their own perspective and incorporate such algorithms into their systems [62]. Incorporating the knowledge of fields like psychology, privacy, and design, we can careful craft algorithmic solutions to problems, mitigate emergent issues of bias, fairness, and discrimination, and execute thoughtful and novel intervention strategies.

Finally, we advocate that the individuals who are the target of predictions should also be considered when developing these algorithms. We especially advocate for participatory approaches of individuals through as focus groups, interviews, and design workshops to better understand their needs, opinions, and interest in this research. As they are both the providers and the recipients of the algorithmic assessments of mental health status, researchers have an obligation to involve them in these decision-making processes. This work is beginning through interview studies [59], and we push researchers to provide future work in this area.

5.2 Developing Best Practices for Methods

> In published work, researchers should disclose study design and methods decisions to promote reproducibility, and the field should agree on what best practices are.

The speed of advancement in this field is impressive – the first papers in this area emerged only five years ago [28, 74]. However, as we note throughout this taxonomy, there are divergent methodological criteria for study design, methods, data privacy, reproducibility, and ethics. How can we understand what these standards are, and arrive at consensus on appropriate methods and protections for research in this area?

To know how to resolve gaps and divergences, the field must know where those gaps are. One method to do so is reflective meta-analyses, reviews, and summative pieces that illuminate the field. We envision such work to be illustrative of both the existing strengths of the fields, and areas for improvement. Systematic literature reviews and recommendations are beginning to be published [91, 5, 18]; in fact, the taxonomy we present here was motivated in part because of this goal. This includes knowledge of end-to-end research design decisions, such as data collection and sampling strategies, issues of consent and privacy management, feature engineering and design, and algorithmic interpretation. We strongly believe that that more meta-work is necessary to document and precisely identify these inconsistencies and gaps. Finally, a benefit of these meta-works and resulting alignments of methodologies is that it enhances replicability of our work in the community.

However, best practices from meta-reviews and analyses must be tempered by careful consideration of advancement in the field as well as respect for individuals as the primary contributors of data and beneficiaries of these systems. Many papers already carefully document their recruitment and consent strategies [26, 68], privacy protections [28, 73], and details on methods and limitations [21, 19]. In addition, consortia such as PERVADE (Pervasive Data Ethics for Computational Research: https://pervade.umd.edu/) and CORE (Connected and Open Research Ethics: https://thecore.ucsd.edu/) offer guidance on how existing ethical codes should be adapted for computational research with sensitive data. We encourage the community to use these as models for best practices in disclosure and transparency into algorithmic design and research.

5.3 Beyond Ethics Boards

> Consider and discuss the implications of this research, outside of the normal considerations of ethics committees. Incorporate ethics as a key value in the research process from the beginning.

The combinations of benign streams of public data into high-accuracy predictions of mental health status creates complex intersections of research outcomes and stakeholders. Fundamentally, this research is human-centered in that the predictions we make are *on people's data*, not on data as an abstracted notion. We draw on the idea of "ethics as a value" in research production, as science in this area has direct implications on people and on society and should be built into the research process. We call researchers to consider the ethics throughout the research process, rather than an afterthought when writing up publications.

When conducting work with direct ties to individuals, we cannot ignore considering implications of this research, even those that extend beyond the purview of ethics boards and oversight committees. Rather than provide checklists for practitioners, we encourage researchers conducting this research to consider and disclose the potentials for benefit and harm. Numerous ethics researchers have cautioned transforming an ethical and sound approach to research into check lists [12, 58]. In particular, Carpenter and Dittrich argue that, by relying on any one piece of ethical guidance, be that an ethics board or a list of best practices, we defer responsibility from considering the risks of a project onto those institutions [12]. We encourage practitioners to be transparent about implications of research in publications, no matter the contribution – a provocative position endorsed by ACM's Future of Computing Academy [42].

6 CONCLUSION

Social media provides a unique perspective into individuals' behaviors and moods. In this paper, we discussed emerging research in using social media data to predict an individual's mental health state. We covered the state-of-the-art in the field and discussed three areas of ethical tension. We offer calls to action to begin to solve these pressing issues, in part because of our belief that this technology can be immensely beneficial in predicting and assessing mental health. We hope that interdisciplinary researchers act on these ideas, and begin to work on solving these pressing challenges in methods, ethics, privacy, and consent.

7 ACKNOWLEDGMENTS

Chancellor and De Choudhury were in part supported by an NIH grant #R01GM112697.

REFERENCES

[1] American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.
[2] John W. Ayers, Theodore L. Caputi, Camille Nebeker, and Mark Dredze. 2018. Don't quote me: reverse identification of research participants in social media studies. *npj Digital Medicine*, 1, 1, 30. DOI: 10.1038/s41746-018-0036-2.
[3] Mitja D Back, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. 2010. Facebook profiles reflect actual personality, not self-idealization. *Psychological science*, 21, 3, 372–374.
[4] Joshua Barrie. 2014. People are freaking out over this new anti-suicide twitter app. (Nov. 2014). https://www.businessinsider.com/people-freaking-out-over-samaritans-twitter-app-2014-11.
[5] 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*, 94–102.
[6] 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.
[7] Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: twitter sentiment and socio-economic phenomena. In *ICWSM*.
[8] boyd, danah and Kate Crawford. 2012. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15, 5, 662–679.
[9] Amy Bruckman. 2002. Studying the amateur artist: a perspective on disguising data collected in human subjects research on the internet. *Ethics and Information Technology*, 4, 3, 217–231.
[10] Jenna Burrell. 2016. How the machine 'thinks': understanding opacity in machine learning algorithms. *Big Data & Society*, 3, 1, 2053951715622512.
[11] Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23, 5, 649–685.
[12] Katherine Carpenter and David Dittrich. 2011. Bridging the distance: removing the technology buffer and seeking consistent ethical analysis in computer security research. In *1st International Digital Ethics Symposium. Loyola University Chicago Center for Digital Ethics and Policy*.
[13] Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *CSCW*. ACM.
[14] Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. 2016. Recovery amid pro-anorexia: analysis of recovery in social media. In *CHI*. ACM.
[15] Qijin Cheng, Tim Mh H Li, Chi-Leung Leung Kwok, Tingshao Zhu, and Paul Sf F Yip. 2017. Assessing suicide risk and emotional distress in Chinese social media: A text mining and machine learning study. *Journal of Medical Internet Research*, 19, 7, (July 2017), 1–10. ISSN: 14388871.
[16] Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication*, 1, 343–359.
[17] Mike Conway. 2014. Ethical issues in using twitter for public health surveillance and research: Developing a taxonomy of ethical concepts from the research literature. *Journal of Medical Internet Research*, 16, 12.
[18] Mike Conway and Daniel O'Connor. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology*, 9, 77–82.
[19] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *CLPsych*, 51–60.
[20] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From adhd to sad: analyzing the language of mental health on twitter through self-reported diagnoses. In *CLPsych*.
[21] Glen Coppersmith, Craig Harman, and Mark H Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *ICWSM*.
[22] Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *CLPsych*.
[23] Patrick Corrigan. 2004. How stigma interferes with mental health care. *American psychologist*, 59, 7, 614.
[24] Amit Datta, Anupam Datta, Jael Makagon, Deirdre K Mulligan, and Michael Carl Tschantz. 2018. Discrimination in online advertising: a multidisciplinary inquiry. In *Conference on Fairness, Accountability and Transparency*, 20–34.
[25] Munmun De Choudhury. 2015. Anorexia on tumblr: a characterization study. In *DH*. ACM, 43–50.
[26] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *CHI*. ACM.
[27] Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *CSCW*. ACM, 626–638.
[28] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *ICWSM*.
[29] Munmun De Choudhury and Emre Kıcıman. 2017. The language of social support in social media and its effect on suicidal ideation risk. In *ICWSM*.
[30] Munmun De Choudhury, Emre Kıcıman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *CHI*. ACM, 2098–2110.
[31] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: hedonometrics and twitter. *PloS one*, 6, 12.
[32] Ezekiel J Emanuel, David Wendler, and Christine Grady. 2000. What makes clinical research ethical? *Jama*, 283, 20, 2701–2711.
[33] 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.

[34] Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4, 1.

[35] Tarleton Gillespie and Nick Seaver. 2016. Critical algorithm studies: a reading list. *Social Media Collective*.

[36] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457, 7232, 1012.

[37] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7, 45141.

[38] Scott A Golder and Michael W Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333, 6051, 1878–1881.

[39] Li Guan, Bibo Hao, Qijin Cheng, Paul SF F Yip, and Tingshao Zhu. 2015. Identifying Chinese Microblog Users With High Suicide Probability Using Internet-Based Profile and Linguistic Features: Classification Model. *JMIR Mental Health*.

[40] Eszter Hargittai. 2015. Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The ANNALS of the American Academy of Political and Social Science*, May, 63–76.

[41] Department of Health and Human Services. 2018. Vulnerable populations. (2018). https://www.hhs.gov/ohrp/regulations-and-policy/guidance/vulnerable-populations/index.html.

[42] Brent Hecht et al. 2018. It's time to do something: mitigating the negative impacts of computing through a change to the peer review process. (Mar. 2018). https://acm-fca.org/2018/03/29/negativeimpacts/.

[43] Jake M Hofman, Amit Sharma, and Duncan J Watts. 2017. Prediction and explanation in social systems. *Science*, 355, 6324, 486–488.

[44] Christopher Homan, Ravdeep Johar, Tong Liu, Megan Lytle, Vincent Silenzio, and Cecilia Ovesdotter Alm. 2014. Toward macro-insights for suicide prevention: analyzing fine-grained distress at scale. In *CLPsych*, 107–117.

[45] Eric Horvitz and Deirdre Mulligan. 2015. Data, privacy, and the greater good. *Science*, 349, 6245, 253–255.

[46] Xiaolei Huang, Lei Zhang, David Chiu, Tianli Liu, Xin Li, and Tingshao Zhu. 2014. Detecting Suicidal Ideation in Chinese Microblogs with Psychological Lexicons. *IEEE UIC-ATC-ScalCom*, 2014, 844–849.

[47] James M Hudson and Amy Bruckman. 2004. "go away": participant objections to being studied and the ethics of chatroom research. *The Information Society*, 20, 2, 127–139.

[48] 2018. International compilation of human research standards. (2018). https://www.hhs.gov/ohrp/international/compilation-human-research-standards/index.html.

[49] Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. 2017. Monitoring tweets for depression to detect at-risk users. In *CLPsych*, 32–40.

[50] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The quarterly journal of economics*, 133, 1, 237–293.

[51] Adam Kramer. 2010. An unobtrusive behavioral model of gross national happiness. In *CHI*. ACM, 287–290.

[52] Adam 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*.

[53] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343, 6167, 1203–1205.

[54] Dave Lee. 2014. Samaritans pulls 'suicide watch' radar app. (Nov. 2014). http://www.bbc.com/news/technology-29962199.

[55] Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. 2014. User-level psychological stress detection from social media using deep neural network. In *MM* (MM '14). New York, NY, USA.

[56] Lydia Manikonda and Munmun De Choudhury. 2017. Modeling and understanding visual attributes of mental health disclosures in social media. In *CHI*.

[57] National Institute of Mental Health. 2018. Eating disorders. (2018). https://www.nimh.nih.gov/health/statistics/eating-disorders.shtml.

[58] Jacob Metcalf and Kate Crawford. 2016. Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society*, 3, 1, 205395171665021.

[59] 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, 22.

[60] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *CLPsych*, 11–20.

[61] Wendy Moncur. 2013. The emotional wellbeing of researchers: considerations for practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1883–1890.

[62] Dan Muriello, Lizzy Donahue, Danny Ben-David, Umut Ozertem, and Reshef Shilon. 2018. Under the hood: suicide prevention tools powered by ai. (Feb. 2018). https://code.fb.com/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/.

[63] Tetsuaki Nakamura, Kay Kubo, Yasuyuki Usuda, and Eiji Aramaki. 2014. Defining patients with depressive disorder by using textual information. *AAAI*.

[64] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Wash. L. Rev.*, 79, 119.

[65] Christopher Norval and Tristan Henderson. 2017. Contextual consent: ethical mining of social media for health research. In *Proceedings of the WSDM 2017 Workshop on Mining Online Health Reports*.

[66] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2016. Social data: biases, methodological pitfalls, and ethical boundaries.

[67] World Health Organization. 2017. Depression and other common mental disorders: global health estimates.

[68] Sungkyu Park, Sang Won Lee, Jinah Kwak, Meeyoung Cha, and Bumseok Jeong. 2013. Activities on facebook reveal the depressive state of users. *Journal of medical Internet research*, 15, 10.

[69] Michael J Paul and Mark Dredze. 2017. Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 9, 5, 1–183.

[70] Michael J Paul and Mark Dredze. 2011. You are what you tweet: analyzing twitter for public health. In *ICWSM*. Vol. 20, 265–272.

[71] James W Pennebaker. 1997. Writing about emotional experiences as a therapeutic process. *Psychological science*, 8, 3, 162–166.

[72] National Commission for the Protection of Human Subjects of Biomedicaland Behavioral Research. 1978. *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Superintendent of Documents.

[73] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6, 1, 15.

[74] Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. *EMNLP*, October, 1348–1353.

[75] Guy Rosen. 2017. Getting our community help in real time. (Nov. 2017). https://newsroom.fb.com/news/2017/11/getting-our-community-help-in-real-time/.

[76] Adam Sadilek, Henry A Kautz, and Vincent Silenzio. 2012. Modeling spread of disease from social interactions. In *ICWSM*, 322–329.

[77] Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D Abowd, and Munmun De Choudhury. 2017. Inferring mood instability on social media by leveraging ecological momentary assessments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1, 3, 95.

[78] H Andrew Schwartz, Johannes Eichstaedt, Margaret L Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards assessing changes in degree of depression through facebook. In *CLPsych*.

[79] H Andrew Schwartz et al. 2013. Characterizing geographic variation in well-being using tweets. In *ICWSM*, 583–591.

[80] H Andrew Schwartz et al. 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PloS one*, 8, 9, e73791.

[81] Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through reddit. In *CLPsych*, 58–65.

[82] T Simms, C Ramstedt, M Rich, M Richards, T Martinez, and C Giraud-Carrier. 2017. Detecting Cognitive Distortions Through Machine Learning Text Analytics. *ICHI*.

[83] 2018. Summary of the hipaa security rule. (2018). https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html.

[84] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *CHI*. ACM, 3187–3196.

[85] Effy Vayena, Marcel Salathé, Lawrence C Madoff, and John S Brownstein. 2015. Ethical challenges of big data in public health. *PLoS computational biology*, 11, 2, e1003904.

[86] James Vincent. 2017. Facebook is using ai to spot users with suicidal thoughts and send them help. (Nov. 2017). https://www.theverge.com/2017/11/28/16709224/facebook-suicidal-thoughts-ai-help.

[87] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2017. Detecting and Characterizing Eating-Disorder Communities on Social Media. In *WSDM*. ACM, New York, NY, USA, 91–100.

[88] Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. 2013. A depression detection model based on sentiment analysis in micro-blog social network. In *PAKDD*. Vol. 7867 LNAI, 201–213.

[89] Yilin Wang, Jiliang Tang, Jundong Li, Baoxin Li, Yali Wan, Clayton Mellina, Neil O'Hare, and Yi Chang. 2017. Understanding and discovering deliberate self-harm content in social media. In *WWW*, 93–102.

[90] Brooke Foucault Welles. 2014. On minorities and outliers: The case for making Big Data small. *Big Data & Society*, 1, 1, 205395171454061.

[91] Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. 2017. Researching mental health disorders in the era of social media: systematic review. *Journal of Medical Internet Research*, 19, 6.

[92] Michael Zimmer. 2018. Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity. *Social Media + Society*, 4, 2.

[93] Michael Zimmer. 2010. "but the data is already public": on the ethics of research in facebook. *Ethics and information technology*, 12, 4, 313–325.

[94] Michael Zimmer and Nicholas John Proferes. 2014. A topology of twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, 66, 3, 250–261.