A Systematic Review of Ethics Disclosures in Predictive Mental Health Research

Leah Ajmani*
ajman004@umn.edu
University of Minnesota
Minneapolis, MN, USA

Stevie Chancellor*
steviec@umn.edu
University of Minnesota
Minneapolis, MN, USA

Bijal Mehta
bm1126@georgetown.edu
Georgetown University
Washington, D.C., USA

Casey Fiesler
casey.fiesler@colorado.edu
University of Colorado Boulder
Boulder, Colorado, USA

Michael Zimmer
michael.zimmer@marquette.edu
Marquette University
Milwaukee, Wisconsin, USA

Munmun De Choudhury
munmund@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

Applied machine learning (ML) has not yet coalesced on standard practices for research ethics. For ML that predicts mental illness using social media data, ambiguous ethical standards can impact peoples' lives because of the area's sensitivity and material consequences on health. Transparency of current ethics practices in research is important to document decision-making and improve research practice. We present a systematic literature review of 129 studies that predict mental illness using social media data and ML, and the ethics disclosures they make in research publications. Rates of disclosure are going up over time, but this trend is slow moving – it will take another eight years for the average paper to have coverage on 75% of studied ethics categories. Certain practices are more readily adopted, or "stickier", over time, though we found prioritization of data-driven disclosures rather than human-centered. These inconsistently reported ethical considerations indicate a gap between what ML ethicists believe *ought* to be and what actually *is* done. We advocate for closing this gap through increased transparency of practice and formal mechanisms to support disclosure.

CCS CONCEPTS

• **Human-centered computing** → *HCI theory, concepts and models*; • **General and reference** → *Surveys and overviews*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

ethics, mental health, systematic literature review, social media

ACM Reference Format:
Leah Ajmani, Stevie Chancellor, Bijal Mehta, Casey Fiesler, Michael Zimmer, and Munmun De Choudhury. 2023. A Systematic Review of Ethics Disclosures in Predictive Mental Health Research. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3593013.3594082

*Both authors contributed equally to the paper

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594082

1 INTRODUCTION

Mental illness is a pressing global health crisis [69], with nearly one in eight people worldwide living with a mental disorder [70]. This urgent issue has inspired researchers to identify innovative avenues for mental health prevention and treatment. As individuals turn to online social platforms to find and share their experiences with mental illness [27], researchers are exploring how and whether this data can identify mental illness with machine learning (ML) and artificial intelligence (AI) technologies. This area has grown into a thriving field of study since its origin in the early 2010s [28, 42]. The state-of-the-art can predict illnesses such as depression [29, 95, 109], anxiety [41, 90, 92], and suicidality [24, 59, 94] with social media data, as well as related experiences such as stress [83]. The impacts of this research could meaningfully change global mental health through earlier identification of illness, supporting quicker and more holistic triage in healthcare, and, ideally, informing digitally-delivered interventions [30].

As the field has matured, there are growing concerns about the research ethics of applied ML systems for predicting mental illness. Mental health prediction creates a ripe environment for possible harm. The domain uses *sensitive* data from a *vulnerable* population to communicate a possible *medical* diagnosis. This research often leverages public online communities, such as Reddit [71] or Instagram [78], but the data collected can include personal information [22, 33]. Experts have raised further concerns about issues around problem framing [15]; methods, such as data handling to maintain data subject privacy [114]; and the need for normative ethical research oversight from ethics review boards [3]. Inadvertent decisions or errors in this space can affect someone's life, such as publicizing an individual's mental health state by using their searchable social media data in research publications [2, 76].

Previous work has suggested what we *ought* to do for more ethical research [3, 16, 101], but it is unclear whether these ethical principles have been actualized in mental health ML research and applications. Sisk et al [97] describe this phenomenon as the "Ought-Is" problem: the challenging gap between what ought to be

and what is, the distance between normative principles and operationalizable practices.[1] An essential step to bridge this gap is to encourage the transparency of practitioners on what *is* happening. Transparency is the crux of applied ML methodological innovation, as it improves the replicability of science and ethics engagement in research [44]. In FAccT, there have been numerous artifacts to increase transparency for methods and datasets [39, 64, 111], yet very little work examining the practice of research ethics through the lens of transparency.

In this paper, we bridge the "Ought-Is" gap with the first step in transparency [44]: documenting the landscape of ethical practices in state-of-the-art research. We conduct a systematic review of ethics disclosures in publications that use social media data to predict mental health states. Building off the methods of a prior systematic literature review [17], we identified 129 relevant papers published between 2013 and 2022. We analyze this dataset using a deductively designed rubric of 13 ethical practices in the field, reporting on human-centered and data-centered ethics practices, and considerations of harm and impact.

Our findings show disparities in disclosure practices–although many practices are being disclosed more over time, others are not. Authors engage in data-centered disclosures that assist in ML replicability over human-centered concerns that impact account holders in datasets. At current disclosure rates, it will take another eight years for all papers on average to disclose 75% of our tracked practices by 2031. This means that, from the origin of the dataset in 2013, it will take nearly two decades for ethics disclosures to coalesce. Moreover, some practices are "sticky" over time, demonstrating the adoption of certain disclosure practices in our dataset.

The encouraging results we did find are slow and inconsistent across disclosure topics, which have been advocated in ethics research [3, 16, 110]. Our work contributes insights into the dynamics of the "Ought-Is" gap in ethics disclosure practices. We underpin the need for researchers to report more transparently on ethical decisions for applied ML systems in mental health. We advocate for more explicit normative standards of ethics and care in computational mental health research. We envision that this increased transparency will empower members of the research community to coalesce on best practices and protect the vulnerable populations that we intend to serve with our technologies.

2 RELATED WORK

2.1 Research Ethics on Social Media and Algorithms

Metcalf and Crawford contend that data "fundamentally changes our understanding of research data to be...infinitely connectable, indefinitely repurposable, continuously updatable and easily removed from the context of collection" [61]. Scholars have argued to expand research ethics to address the gaps left by big data and algorithms in current practice, whether that be related to user privacy [114], data attribution [10], or informed consent [49]. In this section, we overview prior work in research ethics across social media and similar large datasets and algorithmic and AI ethics.

[1]Not to be confused with Hume's "Is-Ought" gap.

2.1.1 Social Media and Large Datasets. There has been increasing discussion on the use of social media data for research. People who create online content, such as tweets and blog posts, often do not know that this content might be used for research purposes [35], especially in relation to mental health [62]. However, this data is increasingly being mined for social media research. Traditional regulatory bodies such as IRBs in the United States and many ERBs in other countries do not consider retrospective analyses of public social media data to be in their scope, as it does not fit the conventional definition of human subjects research [105]. Correspondingly, researchers often do not have clear standards about *how* to proceed with doing this research ethically [106], a problem that is exacerbated by the typical lack of oversight by traditional research ethics review bodies [7].

Social media and the large datasets that are mined from it are used across many research areas; this has provoked many discussions about online communities, big data, and social media [8, 61, 74, 106]. As such, scholars have argued for expanding the scope of social media ethics to address specific practices, whether that be related to user privacy [47], data attribution [10], informed consent [48, 49]. For public health in particular, researchers have honed in on the implications and ethical tensions of using digital data to make research contributions. For example, Vayena et al discuss the ethical obligations of researchers engaging in "digital epidemiology" using public social media data [104].

2.1.2 Applied Algorithm and AI Ethics. Paralleling our interest in social media ethics is the growing interest in the ethical development and use of algorithms and AI to solve social problems. This work has covered wide swathes of work, where some point out low-level challenges [66] and others articulate global principles [52]. The focus on AI has been a more recent trend, as researchers have raised growing concerns about the ethics of algorithms that influence our lives [8]. Critical areas of focus in AI ethics include fairness, justice, equity, and other value-ladened decision-making. Likewise, there has been interest in bridging this ethics gap with human-centered or human-AI interaction work [14].

2.2 Pragmatic Interventions to Support More Ethical Decision-Making

In addition to theorizing about appropriate ethical conduct, researchers have also examined pragmatic strategies to enact ethics in AI, social media, and big data. Sisk et al [97] describe this as the "Ought-Is" problem: the challenging gap between what ought to be and what is. This gap is particularly salient in predictive mental health using large amounts of social media data, where ethical issues and decision-making abound [3, 16, 101].

One common theme across many ethics interventions is increased transparency, primarily through encouraging more documentation and reporting. One common intervention is adapting the peer-review process to include a work's ethics and impact [46]. In machine learning specifically, NeurIPS recently started requiring submissions to include broader impact statements. However, it is unclear whether these requirements for transparency led to more constructive and meaningful engagement [67]. Other scholars have proposed artifact-based interventions to increase transparency, such as model cards [64], datasheets [39], and explainability fact

sheets [98]. However, these artifacts can be difficult to apply – developers struggle to make accurate privacy nutrition labels despite their prominence in research [55].

While these artifacts are valuable frameworks that help people categorize and understand AI [9], they often focus on methodological details or oversight that can be tracked in the moment, such as the provenance of a dataset. Documentation does not necessarily focus on ethics, and discussion of research ethics considerations may be absent or perfunctory (e.g., only a mention of the "publicness" of data) [75]. Scholars note that more open discussion of ethical issues towards creating community norms is important [11], as is evaluating the state of the field as it is [17] to bridge between appropriate theory and praxis. Our review is motivated by these measures towards transparent reporting by focusing on ethical *disclosures* of practice in papers when they do occur. In essence, we seek to understand whether there are implicit standards that guide what researchers choose to report on in their publications.

2.3 Case Area: Mental Health Prediction on Social Media Data

Since the early 2010s, researchers and practitioners have used text and behavioral cues from social media to understand mood, feelings, emotions, and well-being – this work has evolved to a rich sub-field that predicts mental illness in individuals using social media data [16]. Initial work in the area addressed predicting depression [28, 72] and suicidality [60]. The area has now expanded to many other disorders, such as post-traumatic stress disorder [23], schizophrenia [6, 63], eating disorders [19], and anxiety [91]. This also includes symptoms of disorders such as stress [57] and new method adaptations, like incorporating image data [79].

Given the sensitivity of this topic and the inherent ethical tensions in this space, scholars have also written extensively about principles of ethics in the space of health, digital data, and computation. For digital health, this prior work includes meta-analysis pieces of studying non-clinical texts [12], mHealth interventions [100], affective disorders [84], and those that examine social media more specifically [90]. More closely related to our area, Wongkoblap et al had an early work exploring data mining and well-being research on social media, including categories of subjective well-being, happiness, and mental disorders [110]. Guntuku et al conducted an integrative review of 12 recent and famous papers within the space of predicting mental health [42]. They hand-select 12 papers instrumental to the space and examine their labeling schema and methodologies. For ethics, several researchers have identified gaps in public health research using social media data [3, 20]. Chancellor and De Choudhury [17] also studied the methods and algorithmic practices within the field. Closest to our work, Thieme et al [101] found a lack of ethical consideration and user-centered practices amongst HCI literature to support mental health ML applications and called for more ethical practices in the field.

Our review builds on this work by formalizing a process to consider ethics disclosures and practices in this specific field. We examine a large corpus of archived papers in this space to study disclosure patterns. This allows us to close the "Ought-Is" gap between ethics principles and prescriptions and examine what "is" happening in the field as a whole.

3 METHODS

3.1 Systematic Literature Review

Our work uses a systematic literature review (SLR) to identify research articles on mental health prediction using ML and social media data. SLRs allow us to rigorously gather and identify papers in a topic area for scientific scrutiny. Specifically, we expanded on the Chancellor and De Choudhury corpus [17], who are authors of this paper. They conducted an SLR to describe and critique methodological decisions for mental health prediction using social media data. While other reviews are broader (considering mental health and AI generally) [101], Chancellor and De Choudhury's review complements ours by focusing on the same sub-area of research. Second, their review focuses on archived papers. This means that all research is peer-reviewed and the final-stage versions of scientific journal or conference papers. We believe this is the fairest comparison for understanding ethical practices in comparison to posters or late-breaking work, which may be deliberately short while the research is still formative.

To briefly summarize the methods, Chancellor and De Choudhury identified 41 interdisciplinary conferences, journals, and archival workshops where research on mental health and ML may be published [17].[2] They then designed 16 keyword pairs related to mental illness and social media and searched the proceedings and journal records for papers. Using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to guide their filtering and candidate selection process [56], they manually identified articles related to the topic area. Finally, they systematically snowballed through the original corpora's Related Work sections to find other papers about the topic that their initial review had missed, using the same filtering criteria as before. This process generated 75 papers from 2013 to 2018 [17]. Adopting their search methodology and filtering process, we surveyed the literature from 2019 to 2022 with the same method, tools, and filtering criteria. This brought the total number of articles to 129, which we will call the dataset in this paper.

3.2 Rubric Design and Analysis

As we considered how to empirically study ethics disclosures, we noted several challenges in reviewing papers. Ethics are often embedded throughout papers, and many venues do not have explicit "ethics disclosure" sections where authors are expected to list all decisions. On review of the dataset, authors may disclose a methods decision about participant recruitment in the "Methods" section that affects both sampling methods and also indicates an ethics practice. Our goal was to take a balanced approach to studying ethics disclosures against a reasonable expectation of what could be disclosed. We considered applying outside standards to our domain, such as IRB or ERB guidelines, or more general ones such as the Association of Internet Researchers (AOIR) guidelines [37]. We quickly realized that these guidelines could miss domain-relevant information to health and social media and may imply standards that are not relevant to the domain. Therefore, we adopted an iterative process to design a rubric to guide deductive analyses of ethics

[2]For the full methods, please see the Methods section of their paper [17].

in the field. This design is similar to Proferes et al. [75], who used a rubric about research practices to study research on Reddit.

We began with a rough categorization of relevant ethical concepts (e.g. "ethics disclosure", "human subjects"), taken loosely from past work on mental health and social media data [3, 12, 16, 110]. We coded papers with these concepts, and iteratively updated concepts to identify emergent ethical discussions, then recoded previously coded papers for these new concepts to ensure consistency. For example, the category of "ethics review discussed" became several distinct categories such as "subject compensation", "inclusion and exclusion criteria for subjects", and "whether consent was sought". We iterated on this process until no new ethics practices emerged. These categories were then converted to a formal rubric to deductively apply to all articles. Table 1 shows the ethical categories we identified. We clustered these categories into three thematic areas: 1) human-centered disclosures about interacting with participants or account holders; 2) data-centered disclosures about methodological and dataset decisions; and 3) ethics reasoning, harms, and broader impacts sections.

For each rubric item, our coding strategy tracked explicit disclosure of ethics practices. Our interest is in evaluating a paper's transparency of practice, not judging whether their ethical decision was appropriate. This is also important because of geographic location differences, differences in research environments, and methodological differences. We designed for each coding category a spectrum of answers to evaluate if the practice was disclosed. We also added several open-ended sections throughout the rubric to allow for memoing of ideas and notetaking about the dataset to emerge. Please see the Appendix for details about the possible categories that are available in the rubric.

3.2.1 Analysis. Three authors coded the dataset of 129 papers with the rubric. The authors initially cross-coded 10 papers together to ensure consistency of coding and use of the rubric items. They then split the dataset, closely read each paper, and coded each paper independently along the final rubric. The three coding authors collaborated on our findings and brought initial insights to three expert coauthors, who provided information about ethics in the domain and research ethics practices in general. Two are experts in social media research ethics and one is an expert in mental health and social media data. The experts reviewed the drafts of the findings and gave recommendations on important emerging considerations in the field that were absent in our initial analysis.

3.2.2 Reflexive Considerations. Given our focus on ethical disclosures, we disclose our position as researchers to help contextualize our findings. As a team, we value human-centered and community-centered approaches to applied ML and AI research. Two authors on this paper have their research included in the dataset, which means they are both objects of the critique in this paper and also critical insiders to this domain [4]. The other authors on this paper are not in the dataset, and help provide important insights that may be missing given our own subjective perspectives. We are cautiously optimistic that applied ML can be carried out ethically.

4 BACKGROUND

4.1 Dataset Overview

Our SLR expansion resulted in 129 articles published between 2013 and 2022. The distribution of publications by year can be found in the Appendix. Given that predictive mental health is an interdisciplinary field, our dataset spans venues such as information retrieval (SIGIR), human-computer interaction (CHI), and digital health (JMIR). We found the most work (13 papers) from CLPsych.

Many social platforms were studied across the dataset. Some papers used pseudonymous platforms, such as Twitter or Reddit, but we also found less anonymous platforms, such as Facebook. Notably, about 36% (47/129) of the articles used secondary data sets that the authors did not collect themselves. Often, these were popular benchmark sets, such as the eRisk [71] or myPersonality datasets [99], used by 8/129 and 3/129 papers, respectively.

4.2 Ethics Disclosure Practices Across Areas

In this section, we qualitatively describe individual disclosures in our dataset that exemplify our rubric (listed in Table 1). We found disclosures varied in length, structure, and section of paper in which they appeared. Some papers had only one short disclosure when describing their methods [53, 54, 82], whereas others had lengthy sections with several paragraphs reasoning about ethics or potential harms of the research [93, 96]. What counted as an ethics disclosure also varied between papers – some authors treat these disclosures as opportunities to reason about more complex ethical issues [40, 43, 108] whereas some took a more list-like format to clearly and quickly disclose practices [81, 112]. Next, we describe emergent patterns around ethics disclosures, clustered into the two largest ethics disclosure groupings in our dataset: human-centered and data-centered disclosures.

4.2.1 Human-Centered Disclosure Practices. Ethics review boards, such as university IRBs or ERBs, oversee human-subjects research, of which public social media research is not. We found that 6/13 of our rubric items aligned with principles stipulated by these human-subjects-driven concerns (see Table 1). Recent work argued collecting social media data implies a level of interaction that warrants similar standards and disclosures [3]. For example, researchers would disclose ethics review board approval by providing IRB reference numbers for their studies [73] or noting the reason why ethics board approval was not sought – *"Approval from the institutional review board was not sought because these data were freely available in the public domain and researchers had no interaction with the users"* [6]. This reasoning is common and is consistent with the typical definition of human subjects research as provided by federal statute 45 CFR §46 in the United States that governs IRBs [36].

4.2.2 Data-Centered Disclosure Practices. We also found five practices around data collection and handling (see Table 1). Some papers disclosed multiple data-handling strategies in specific ethics sections: *"We paraphrased and anonymized all examples...We kept all user data separately on private servers linked to the raw text and accessible only through anonymous IDs"* [85]. More often, we found these disclosures scattered throughout the paper or buried in the

	% Disclosed	Definition
Human-Centered		
Data source (public vs. private)	70.54	Origin of the data from and the visibility of this data to the general public
Interaction with Subjects	53.49	How researchers engaged with the account holders in the dataset, if at all
Ethics board mentioned	38.76	Whether and how an ethics review board was mentioned
Consent	24.03	Whether and how participants consented into the research, or logic about why consent was not sought
Compensation	9.30	Money received for participating in research, if any
Plan to inform subjects/gatekeepers	7.75	Intention to relay research results back to participants or communities
Data-Centered		
Subject inclusion/exclusion criteria	96.12	Standards to include participants in datasets
Modification of personal data	81.40	Concealing identifying information in the paper if used quotes or examples
Data de-identified	33.33	Anonymizing or changing data in the research process
Data sharing protocol	33.33	Guidelines for sharing resources (datasets, models, etc.)
Data storage	9.30	Methods of storing data
Impact		
Harms consideration	50.39	Direct or indirect mention of negative implications or impacts
Ethics consideration section	47.29	Specific written structure that discusses ethics concerns

Table 1: An overview of our rubric. Includes the proportion of papers in our systematic literature review (n = 129) that disclosed ethics practices. See Appendix for more details on how each disclosure was coded.

relevant figure caption [18]. These would occasionally be cross-referenced with the platform's anonymity standards: *"Reddit platform enables free, unobtrusive, and honest sharing of mental health concerns because a patient is completely anonymous"* [38]. However, the anonymity of pseudonymous platforms, such as Reddit, is a contentious topic among researchers [1, 75, 77].

4.2.3 Impact and Harms Considerations. About half of the articles included discussions of research harms, ethics, or broader impacts. One pattern was interleaving harm considerations with methods and limitations. For example, one paper considered the impacts of potential harms caused by the accuracy of the model: *"...the prediction scores of our classification models, even the accurate ones, are not well calibrated and thus are not an accurate uncertainty estimator of mental health risk"* [51]. Others cited privacy concerns when describing their data collection processes: *"Strict anonymity was nearly impossible to guarantee to participants, given that usernames and personal photographs posted to Instagram often contain identifiable features."* [78].

5 FINDINGS

5.1 Disparities in Disclosure Practices

In Table 1, we summarize the 13 individual ethical principles and the number of papers that remark on the subject. Recall from the Methods that our evaluation of disclosure is not judgmental about the quality of ethical reasoning – we simply track if the concept was mentioned to better understand transparency of practice. Therefore, the percentages in this table represent the papers that mentioned the issue at all (either in passing or in-depth), divided by the total number of papers in the dataset (n=129).

There is a large disparity in disclosure rates in the 13 ethics practices. Some disclosure practices are already adopted in the research. Three out of 13 of our ethics disclosure practices were done by 70% or more articles - describing the data source (public vs. private) (70.54%); subject inclusion/exclusion criteria (96.12%); and whether the data was modified in the paper (81.4%). However, other practices were rarely disclosed. Another three of our rubric items were disclosed in <10% of papers: compensation (9.3%), plan to inform subjects (7.75%), and data storage practices (9.3%). This disparity suggests that predictive mental health researchers have not yet identified a shared set of transparency standards despite calls from highly cited prescriptive papers [3, 16, 21, 101].

5.1.1 Ethics practices are more disclosed when they serve ML methodological purposes. Considering Table 1, ethical practices are disclosed when they have a connection to other methodological evaluation metrics, such as reproducibility or rigor. The most frequently disclosed criteria (disclosed in >70% of papers) all relate to methodological details necessary to replicate machine learning experiments. For example, subject inclusion/exclusion criteria is necessary to construct training and test datasets; modifying personal data is a component of pre-processing, and data source information is necessary for replicability and future gathering of said data.

Qualitatively, these often appeared in publications as "implicit" ethics disclosures–practices that were highlighted in a paper's methods but not explicitly called out as ethics concerns. For example,

> In this study, a dataset containing publicly available Reddit posts was used... usernames were not downloaded from the data source during this study, and ethics committee approval was not sought. [32]

This example is taken from the "Data Collection" section and performs several ethics disclosures. This discloses "data source" (public data from Reddit), "data de-identification," and "IRB/ERB

approval". Similarly, papers from our dataset would make methods disclosures to serve goals of documenting pre-processing steps.

These "implicit" disclosures are encouraging; they show the potential for connecting ethics and methods as a mechanism for increasing ethics disclosures. However, relying solely on methods-oriented disclosures may decrease the importance of human-centered ones, which have been prescribed by ethicists [16, 45, 74] but are more abstract than data-handling disclosures. In other words, these implicit disclosures may be one component of increasing transparency but are not sufficient to reach the field's self-stated goal of increasing human-centered disclosures.

5.1.2 Ethics considerations sections. Next, we consider the *explicit* disclosures that authors made in our dataset. We identified specific sections, typically titled "Ethics Considerations" or "Broader Impacts", where the authors call out their ethical practices, procedures, or decisions. We include both required disclosures by the venue (such as in NeurIPS, ICWSM, or ACL [67]) and sections authors used to draw attention to their decision-making. Despite the increasing number of venue standards, we found large qualitative inconsistencies in these sections across the predictive mental health research community.

Some ethics considerations sections were cursory (no more than a sentence or two) and included procedural reasoning. Some sections, though titled "Ethical Considerations," were brief and did not disclose many of our rubric items: *"The data set and methods used in this work are publicly available and do not involve any ethical or moral issues"* [25]. Other minimalist ethics disclosure sections conflated ethics discussions with ERB review [115]. These findings align with work in other domains that show perfunctory ethics discussions in Reddit research [75] and in the broader impact sections in NeurIPS papers [67].

Many papers had longer sections on ethics disclosure and had more content; nonetheless, there were differences in what constitutes "ethics discussion." For example, some papers used an "Ethics Considerations" section to outline data practices [88]. Meanwhile, others reasoned about potential bad actors: *"From the perspective of privacy concerns, organizations with vested interests (e.g. insurance companies) may be motivated to infer this information automatically"* [40]. Still, others acknowledged the norms of their field and how they either follow or deviate. For example, *"Additionally, in a departure from traditional practices in the NLP community, the data underlying this work will only be shared with researchers who both (1) provide a research design or other public health justification for the use of the data and (2) agree to take the necessary efforts to secure the data"* [108]. Although disparate, these disclosures are promising, as they comply with ethics prescriptions surrounding data-handling [3], broader impact consideration [74], and norm creation [11].

Our results suggest that there is great variety of what practitioners consider explicitly to be "ethics disclosure." While we saw many insightful ethics considerations in our dataset, there were no indicators of a shared understanding of ethics considerations across the field. Particularly in peer-review science, where researchers use publications to learn from one another [75], this variation can cause confusion about what ethics even is, thereby, making the "Ought-Is" gap all the more challenging to overcome [97].

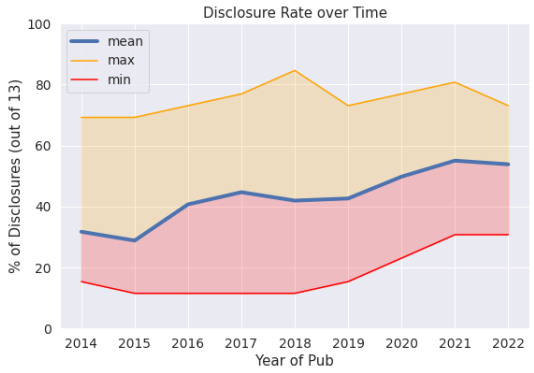


Figure 1: Mean (blue), max (yellow), and min (red) of the percentage of rubric items (out of 13 total) that were disclosed over time. We see a moderate upwards trend in the mean rubric items disclosed.

5.2 Disclosure Trends over Time

Next, we analyze trends in disclosure patterns over time. All line graphs represent counts in a given year, smoothed with a two-year rolling average to account for unusual annual spikes or lags in publication output. Temporally studying our findings allows us to reason about the growth and change patterns in our dataset.

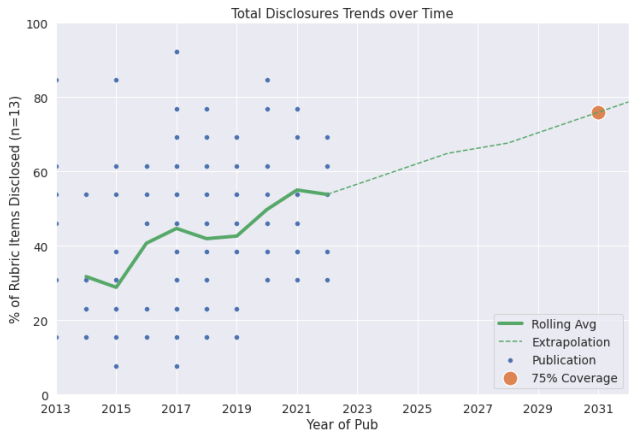


Figure 2: Average disclosures over time. Solid green (2013-2022) represents the data from our systematic review while dashed green is an extrapolation. Based on a rate of change of 2.76%, the field will average disclosing 75% of our rubric items by 2031, shown in orange.

Figure 1 shows the average number of disclosures in a given year, as well as the minimum and maximum number of disclosures. We see an encouraging upward trend in the number of disclosures, indicating that the field is moving towards more disclosure practices. In 2013, papers had an average of 5.33 disclosures. In contrast, in 2021 and 2022, we saw an average of 6.83 disclosures per paper.

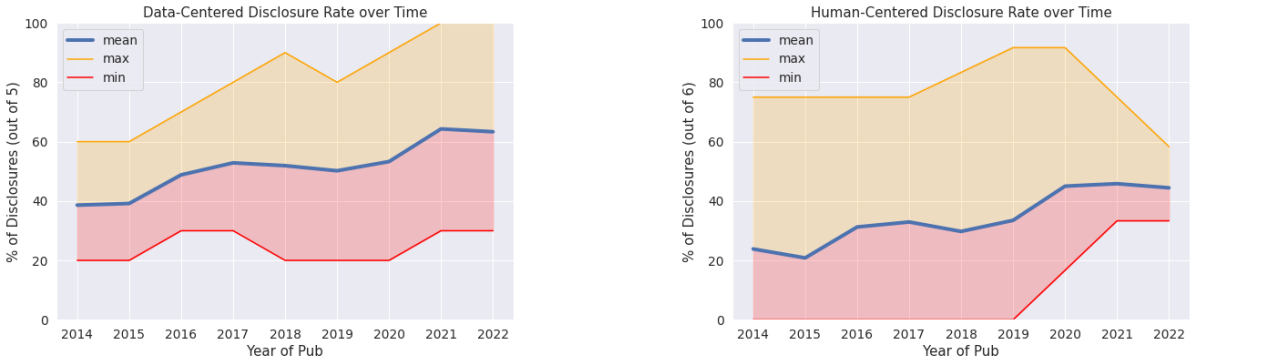


Figure 3: Mean, max, and min trends for data-centered disclosures (left) and human-centered disclosures (right). While there is a general upwards trend in both categories, we find human-centered practices are, on average, less disclosed.

Our results indicate that this trend is slow-moving. Specifically, we find a rate of change of 2.76% in the average number of disclosures ($(rolling\_mean_{2022} - rolling\_mean_{2014})/(2022 - 2014)$). Figure 2 demonstrates how the average disclosure rate would continue to increase if the rate of change stayed at 2.76%. Hypothetically, we would not see 75% coverage until 2031. Said another way, if the current ethics disclosure trends continue, it will take eight years for the field's average to exceed 9/13 of our rubric items. Contextualized from the start of the dataset (2013), this means that the average pattern of disclosing ethics practices could take 18 years, or nearly two decades, to reach 75% coverage. Given the rapid rate of publication and innovation in machine learning for mental health [17], this suggests that the field may need ethics disclosure interventions to accelerate the slow growth we found in our dataset, even before normatively deciding what is appropriate to do in situations.

We find similar trends when broken down by disclosure topics. Figure 3 shows disclosure rates over time by data-centered and human subject-centered disclosures. We see encouraging and increasing trends for both kinds of ethics disclosures. However, the average human-centered disclosures never exceeded 50% of our rubric items. We also note that the maximum, or "ceiling", of human-subjects disclosures is decreasing over time ($max_{2013} = 83.33\%; max_{2022} = 50\%$). This is surprising given the amount of conversation surrounding human-subjects concerns in research [35] and the rise of ethics prescriptions calling for the application of more stringent ethics standards to social media data [3, 110].

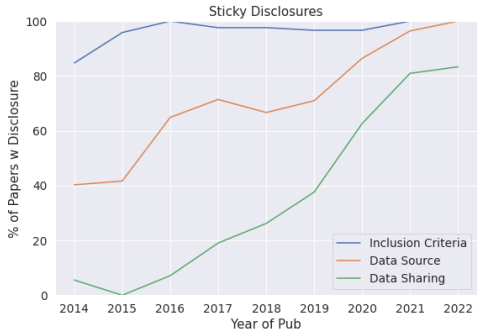
In summary, our dataset highlights the growth in ethics disclosure practices. However, we find that this growth is slow and, therefore, may require accelerators, such as transparent guidelines and disclosure artifacts. Moreover, the current growth patterns are not equally "lifting" all disclosures. Future interventions would need to specifically focus on human-centered concerns, which are particularly under-disclosed in our dataset.

5.2.1 Stickiness. Next, we examine how individual practices of disclosure are more consistently disclosed over time. We define the "stickiness" of disclosure patterns, such that a disclosure practice is sticky if: 1) the practice has an increasing trajectory to high adoption (>66% of yearly papers by 2021) and 2) the disclosure percentage of this practice remains stable (it does not dramatically dip). Stickiness highlights whether a disclosure practice is becoming an implicit norm in the field–an unwritten but widely accepted rule [31].

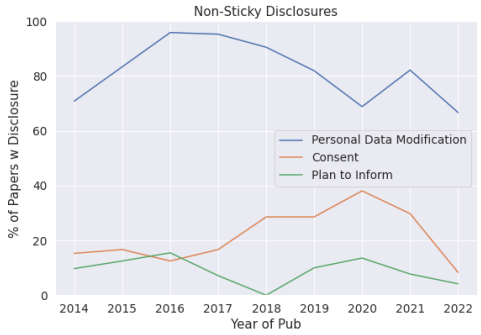
We find that different types of ethics disclosures have differing "stickiness." In Figure 4, we show the disclosure rates across six of the individual ethics practices. We first note a promising trend – some practices had a consistent upward trend of being disclosed over time, thereby becoming "sticky" in the area. Three practices (data source; subject inclusion criteria; and data sharing protocols) are practices that have either stayed consistent or gotten more popular over time (Figure 4a). The disclosure of data-sharing protocols increased from 11% of papers in 2013 to 83% of papers in 2022. Likewise, disclosing subject inclusion/exclusion criteria has been somewhat sticky from the start (disclosed in >75% of papers) from 2013-2022, but has remained consistent in the dataset. This consistent growth suggests that disclosure of data-sharing practices stuck with time. We hypothesize that, for these findings, stickiness is partly due to increased methodological considerations in applied ML, such as reproducibility (see prior Findings).

However, not all disclosure practices had the same stickiness, suggesting that long-term adoption of certain ethics discussions may be more difficult. In Figure 4b we see that consent and plans to inform subjects never became prevalent in our dataset. Plans to consult with or share results back to the individuals or communities were disclosed in 7.75% of papers in our dataset and, as the line graph shows, informing subjects has not picked up to become a consistent practice. This is in notable contrast to work that has found informing data subjects as an important factor in their perceptions of research ethics [35]. Among papers that used personal data, such as tweets or Instagram posts, disclosing how that data was modified did not consistently trend upwards over time. Modification disclosures were present in 66% of papers in 2013, but in 33% of papers in 2022. Because stickiness is a function of adoption, these results suggest that some practices are more difficult for the field to organically and consistently disclose.

In addition to the quantitative trends we saw, we also found verbiage and reasoning that was qualitatively sticky. For example, of the 36% of papers that used secondary datasets, many simply pointed to the benchmark set and omitted usage-dependent disclosures, such as how they chose to store, modify, or share the data – *"Pirina and Çöltekin built a data set for depression detection*



(a) Trends over time for subject inclusion/exclusion criteria, data source description, and data sharing protocols.



(b) Trends over time for personal data modification, consent reasoning, and plan to inform subjects or community gatekeepers.

Figure 4: Trends over time for stickiness

based on Reddit, which was named the Reddit data set. The samples in the Reddit data set are collected from the Reddit platform" [80]. We found this to be especially salient in papers that used the eRisk dataset, which was first released in 2017 and was used throughout our dataset. Many papers linked to the dataset, but there was no indication of proper use for the dataset or how it was de-identified. Trotzek et al. [103] were the only authors to note a privacy issue in the dataset itself:

> Since users for the control group were collected by selecting users that had posted recently when the dataset was collected...the timestamps also contain a hidden feature that could be exploited..all models created for this paper completely discard the timestamp information and a detailed analysis of this fact has been sent to the organizers of eRisk to prevent this in future tasks [103]

We found a similar trend in reasoning about consent being qualitatively sticky. Several papers argued that asking account holders on public social media websites to participate in their research could be coercive. These arguments start in 2015 and appear again through 2022 [68, 85–89]. It is important to note that consent is typically considered an anti-coercion mechanism because it gives people the choice to participate in research [35, 50]. We could not locate additional justification for this argument and, therefore, these papers run against common knowledge of consent in research ethics and recommendations in the space [3, 16, 45, 110]. This perspective warrants discussion because it could be unintentionally propagated through the dataset despite its misalignment with widely accepted principles around consent. However, we are concerned that these practices could become sticky if prior work cites back to these examples as motivation for future methods decision-making.

To conclude, when evaluating how individual disclosures trend over time we find variability in a disclosure's stickiness. The existence of sticky rubric items is encouraging; it suggests that disclosures can eventually grow into common practices. However, we find that less sticky disclosures are common themes in influential ethics papers on the topic area, such as consent. This leads us to believe that these less sticky practices are more difficult to discuss and, thereby, disclose. Qualitatively, we find that ideologies around benchmark datasets and consent were propagated throughout our dataset. This suggests that stickiness has the power to unintentionally create an implicit norm.

6 DISCUSSION

6.1 Sticky Ethics Practices and Norm-Setting in Research

Our analysis demonstrates "stickiness", that certain ethical practices become more consistently and regularly practiced over time. Our results indicate promising results about stickiness for ethical disclosures in our dataset - there are several categories of disclosures happening more frequently as time goes on. While only 11% of papers in our dataset discussed data-sharing practices in 2013, this number steadily increased to 83% by 2022.

Sticky ethics practices are evidence that implicit norm maintenance and creation can happen in a research area through peer-reviewed publications. In policy-based disciplines, such as laws and economics, stickiness is a necessary component to creating implicit norms–"unwritten" best practices that are typically enforced by peers [13, 26, 31, 34]. For ethics, implicit norms can create moral standards at little to no cost [31] and imply that researchers are learning from one another to make cooperative progress. Moreover, disciplines such as medicine rely in part on strong normative practices, captured in notions of duty and obligation [65].

However, implicit norms for ethical disclosures are not a perfect system for maintaining high-quality ethical practices. One obvious challenge with implicit norms is their implicitness – meaning that to inspect normative practices in a field, analyses like ours need to be conducted to understand what is going on. In addition, sociologists and ethicists have found that implicit norms can allow undesired practices to go unchecked [5]. In our dataset, we found this when "sticky practices" about consent being coercive were repeated in papers without strong evidence supporting it. In contrast, prescriptive papers have heavily discussed consent and its importance in machine learning applications that use social media data [16, 101, 113]. In this case, an implicit acceptance of consent (through a lack of

rigorous engagement because of normative practices in peer review) means that these become "codified" as explicitly permissible. At their worst, implicit norms can unintentionally create a tyranny of the majority [58], and because they are held and imposed by large groups of people, can disproportionately affect minority or marginalized communities [102]. Our results found that implicit norms surrounding ethics disclosures are being created and propagated. While this has the potential to progress the field closer to certain principles, our results suggest that we cannot solely rely on stickiness to propagate norms. In other words, we must bring "unwritten" rules to the written page in order to assess whether the field's practices are moving towards its principles.

6.2 The "Ought-is" Gap

Discussions on ethics principles are especially salient in predictive mental health work, where the data can be sensitive and the subjects may be in emotionally vulnerable states. Recently, there has been a movement of prescriptive work in our domain area on which ethical tensions are particularly salient. In research ethics, these prescriptions are assessments of what "ought" to be: what practices are we, as researchers, morally obligated to engage with [97]. These authors have called for informed consent at scale, fairness measures, and multidisciplinary involvement, to name a few [3, 16, 101, 110]. In fact, many of these prescriptive works have clear and explicit guidance on some of these issues – for example, Benton et al [3] prescribe data handling techniques, such as significantly anonymizing data in quotes and publications.

Our research finds a gap between what is prescribed (i.e., seen as important) and what is being disclosed. This gap is especially salient with regard to consent. We found general plurality in prescriptive work on discussing consent as researchers who use social media data [3, 16, 50, 101, 110]. However, only 24.03% of papers in our dataset remarked on consent practices despite this plurality. Moreover, we found that getting consent was not a "sticky" practice; our dataset did not evidence consent trends substantively changing over time (2013 = 15.27%; 2022 = 8.33%).The principles surrounding consent are not reflected in the disclosure practices authors engage in. Furthermore, authors prioritize disclosures that serve methodological concerns, such as describing the data source (public vs. private) (70.54%); subject inclusion/exclusion criteria (96.12%); and whether the data was modified in the paper (81.4%). While these data-centered disclosures contribute to ethics discussions, many prescriptive papers emphasize human-centered concerns, such as informing community gatekeepers [16], justice for data subjects [101], and consideration of user-interaction [3]. These concerns were disproportionately underrepresented in our dataset, especially in comparison to data-centered ones. For example, only 7.75% of papers discussed plans to inform community members, such as moderators, or data subjects.

We hypothesize several reasons why this transparency gap exists, especially when considering the human and data-centered divisions we note. Many publication venues create constraints that complicate publishing ethics disclosures. For example, conferences and journals may have strict page limits or word counts. This may lead to a tradeoff where researchers prioritize methods or results over ethics discussions because of page limits, making it logistically easier to put in ethical disclosures about methods. Second, authors may be cautious about concerns of "redundancy" and therefore not discuss concerns that are covered by their local ERB. For example, we found papers that stated they avoided any ethics concerns because they were IRB/ERB exempt. This suggests that any future interventions, such as the ones we propose below, must account for these constraints to encourage researchers to disclose important information in their published work.

6.3 Future Work: Pragmatically Closing The "Ought-Is" Gap

Given these challenges, what are the next steps to resolve these gaps? Our results highlight the prevalence of the "Ought-Is" gap in discussions of ethics in predictive mental health research. Notably, we recognize that this gap is not intentional nor malicious. Principles are difficult to operationalize [65, 107] and interventions take time to be effective. However, there are opportunities to improve disclosures, increase transparency, and, eventually, lead to better outcomes for field-wide conversations on appropriate ethics in applied AI in future work. We provide several ideas below:

6.3.1 More Open, Interdisciplinary Conversations About Ethical Practices. Our results indicate that ethics is more than a matter of "yes or no." For example, we found unique insights by qualitatively understanding *why* certain authors were not acquiring consent. This suggests that the field needs spaces for researchers to reason about their decisions before those decisions are judged as "good or bad." Some venues have attempted to create these spaces through town halls on research ethics [11], reflections on broader impact statements [67], or discursive workshops. We are excited by venues like FAccT and CLPsych to allow for these conversations and deliberations to happen.

6.3.2 More Space in Papers For Disclosure and Reasoning. While the conversation is important, incentivizing better practices in peer-review processes is crucial so authors can disclose ethical decision-making and reasoning without worrying they will compromise the methodological rigor of their work. We suspect that authors had to sometimes choose between page limits and including ethics disclosures. We also encourage venues to be more supportive of this by giving dedicated space in the paper for ethics disclosures and reviewers more ability to scrutinize and engage with these disclosures. Additionally, venues could allow for longer appendices so that ethical information is accessible within the paper itself. In tandem with these extra documents, it is important that we include them in the peer-review process. In alignment with previous provocations on incorporating ethics into peer review [46], we suggest that venues empower reviewers to ask for more ethics disclosures or supplementary materials if not included in the paper.

6.3.3 Context Documents to Increase Transparency. Recent work has explored the use of context documents to increase transparency in AI, inform appropriate use, and promote ethics reflections [9, 39, 64]. Specifically, we see the potential for context documents to move beyond documentation and scaffold the disclosure process for researchers. If these documents can become interactive or more useful for researchers, we can empower authors to become more familiar with ethics and tackle challenging problems.

6.4 Limitations

An important limitation is that we do not know what people actually did; rather, we focus on what they disclosed. This may mean ethics practices are in fact happening but are not being logged in research papers, thereby meaning we undercount the rates that the practice happens. We believe this is more plausible for data-driven disclosures (such as storing data on a protected server) than human-centered ones given our dataset. Interviews would be necessary to speak to researchers but may be subject to hindsight bias given that some papers in this dataset are 10 years old.

7 CONCLUSION

In this paper, we take the perspective that transparency is essential for better ethics practices. Our review criteria are motivated by calls for increased transparency about ethics in applied machine learning, and we analyze the ethics practices that authors disclose to promote more disclosure. We find some promising results: overall trends of ethics disclosures are increasing over time. However, this growth is slow and mixed across disclosure categories. Understanding this Ought-Is gap [97] is crucial to finding ways to close it and bring out intentions in alignment with our goals of doing more sensitive and human-centered research.

We propose expansions of current solutions such as town halls, changes to the peer-review process, and context documents to increase transparency. We underpin the need for the field to accelerate and standardize ethics disclosures, especially in mental health research where there is a large potential to cause harm and compromise people's well-being.

ACKNOWLEDGMENTS

We thank Hanlin Li and our reviewers for their invaluable feedback on this paper. De Choudhury was partly funded by National Institute of Mental Health grant R01MH117172. Chancellor completed a portion of this work while at Northwestern University.

REFERENCES

[1] Tawfiq Ammari, Sarita Schoenebeck, and Daniel Romero. 2019. Self-declared Throwaway Accounts on Reddit: How Platform Affordances and Shared Norms enable Parenting Disclosure and Support. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), 1–30.

[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 (2018), 30.

[3] 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.

[4] Roni Berger. 2015. Now I see it, now I don't: Researcher's position and reflexivity in qualitative research. *Qualitative research* 15, 2 (2015), 219–234.

[5] Cristina Bicchieri and Yoshitaka Fukui. 1999. The Great Illusion: Ignorance, Informational Cascades, and the Persistence of Unpopular Norms. In *Experience, Reality, and Scientific Explanation: Essays in Honor of Merrilee and Wesley Salmon*, Maria Carla Galavotti and Alessandro Pagnini (Eds.). Springer Netherlands, Dordrecht, 89–121.

[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 (2017), e7956.

[7] Kathleen M Blee and Ashley Currier. 2011. Ethics beyond the IRB: An introductory essay. *Qual. Sociol.* 34, 3 (2011), 401–413.

[8] Danah Boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15, 5 (2012), 662–679.

[9] Karen L Boyd. 2021. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–27.

[10] 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 (2002), 217–231.

[11] Amy S Bruckman, Casey Fiesler, Jeff Hancock, and Cosmin Munteanu. 2017. CSCW Research Ethics Town Hall: Working Towards Community Norms. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17 Companion)*. Association for Computing Machinery, New York, NY, USA, 113–115.

[12] 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 (2017), 649–685.

[13] Emanuela Carbonara, Francesco Parisi, and Georg von Wangenheim. 2012. Unjust laws and illegal norms. *Int. Rev. Law Econ.* 32, 3 (2012), 285–299.

[14] Stevie Chancellor. 2022. Towards Practices for Human-Centered Machine Learning. *arXiv preprint arXiv:2203.00432* (2022).

[15] Stevie Chancellor, Eric PS 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 (2019), 1–32.

[16] Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB 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*. 79–88.

[17] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine* 3, 1 (2020), 1–11.

[18] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms matter: Contrasting social support around behavior change in online weight loss communities. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2018-April.

[19] Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. 2016. Recovery amid pro-anorexia: Analysis of recovery in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2111–2123.

[20] 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 (2014), e290.

[21] Mike Conway. 2014. Ethical issues in using Twitter for public health surveillance and research: developing a taxonomy of ethical concepts from the research literature. *J. Med. Internet Res.* 16, 12 (2014), e290.

[22] Glen Coppersmith, Mark Dredze, Craig Harman, Holli, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *CLPsych*. 1–10.

[23] 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 *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*. 1–10.

[24] Glen Coppersmith, Anthony Wood, Kim Ngo, Ryan Leary, and Anthony Wood. 2016. Exploratory analysis of social media prior to a suicide attempt. In *CLPsych*. From Linguistic Signal to Clinical Reality. North American Chapter of the Association for Computational Linguistics, San Diego, Califor- nia, USA, In Proceedings of the Workshop on Computational Linguistics and Clini- cal Psychology, 106–117.

[25] B. Cui, J. Wang, H. Lin, Y. Zhang, L. Yang, and B. Xu. 2022. Emotion-Based Reinforcement Attention Network for Depression Detection on Social Media: Algorithm Development and Validation. *JMIR Medical Informatics* 10, 8 (2022).

[26] Martin Davies. 1995. Two Notions of Implicit Rules. *Philos. Perspect.* 9 (1995), 153–183.

[27] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

[28] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

[29] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 2 (2013), 128–137.

[30] Munmun De Choudhury and Emre Kiciman. 2018. Integrating artificial and human intelligence in complex, sensitive problem domains: Experiences from mental health. *AI Magazine* 39, 3 (2018), 69–80.

[31] Robert C Ellickson. 1998. Law and Economics Discovers Social Norms. *J. Legal Stud.* 27, S2 (1998), 537–552.

[32] Ahmet Emre Aladag, Serra Muderrisoglu, Naz Berfu Akbas, Oguzhan Zahmacioglu, and Haluk O Bingol. 2018. Detecting Suicidal Ideation on Forums: Proof-of-Concept Study. *JOURNAL OF MEDICAL INTERNET RESEARCH* 20, 6 (2018).

[33] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–16.

[34] Yuval Feldman and Oren Perez. 2009. How law changes the environmental mind: An experimental study of the effect of legal norms on moral perceptions and civic enforcement. *J. Law Soc.* 36, 4 (2009), 501–535.

[35] Casey Fiesler and Nicholas Proferes. 2018. "Participant" perceptions of Twitter research ethics. *Social Media+ Society* 4, 1 (2018), 2056305118763366.

[36] Office for Human Research Protections (OHRP). 2016. 45 CFR 46.

[37] franzke, aline shakti, Anja Bechmann, Michael Zimmer, Charles Ess, and the Association of Internet Researchers. 2020. Internet Research: Ethical Guidelines 3.0. (2020).

[38] M. Gaur, A. Sheth, U. Kursuncu, R. Daniulaityte, J. Pathak, A. Alambo, and K. Thirunarayan. 2018. "Let me tell you about your mental health!" Contextualized classification of reddit posts to DSM-5 for web-based intervention. In *International Conference on Information and Knowledge Management, Proceedings*. Association for Computing Machinery, 753–762.

[39] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[40] S.C. Guntuku, A. Buffone, K. Jaidka, J.C. Eichstaedt, and L.H. Ungar. 2019. Understanding and measuring psychological stress using social media. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*. Association for the Advancement of Artificial Intelligence, 214–225.

[41] S.C. Guntuku, D. Preotiuc-Pietro, J.C. Eichstaedt, and L.H. Ungar. 2019. What twitter profile and posted images reveal about depression and anxiety. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*. Association for the Advancement of Artificial Intelligence, 236–246.

[42] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (2017), 43–49.

[43] X. Guo, Y. Sun, and S. Vosoughi. 2021. Emotion-based Modeling of Mental Disorders on Social Media. In *ACM International Conference Proceeding Series*. Association for Computing Machinery, 8–16.

[44] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Massive Analysis Quality Control (MAQC) Society Board of Directors Shraddha Thakkar 35 Kusko Rebecca 36 Sansone Susanna-Assunta 37 Tong Weida 35 Wolfinger Russ D. 38 Mason Christopher E. 39 Jones Wendell 40 Dopazo Joaquin 41 Furlanello Cesare 42, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, et al. 2020. Transparency and reproducibility in artificial intelligence. *Nature* 586, 7829 (2020), E14–E16.

[45] Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the State of Social Media Data for Mental Health Research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*. 15–24.

[46] Brent Hecht, Lauren Wilcox, Jeffrey P Bigham, Johannes Schöning, Ehsan Hoque, Jason Ernst, Yonatan Bisk, Luigi De Russis, Lana Yarosh, Bushra Anjum, et al. 2021. It's time to do something: Mitigating the negative impacts of computing through a change to the peer review process. *arXiv preprint arXiv:2112.09544* (2021).

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

[48] 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 (2004), 127–139.

[49] Luke Hutton and Tristan Henderson. 2015. " I didn't sign up for this!": informed consent in social network research. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 9. 178–187.

[50] Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark S Ackerman, and Eric Gilbert. 2021. Yes: Affirmative Consent as a Theoretical Framework for Understanding and Imagining Social Platforms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21, Article 403)*. Association for Computing Machinery, New York, NY, USA, 1–18.

[51] Z. Jiang, S.I. Levitan, J. Zomick, and J. Hirschberg. 2020. Detection of mental health conditions from Reddit via deep contextualized representations. In *EMNLP 2020 - 11th International Workshop on Health Text Mining and Information Analysis, LOUHI 2020, Proceedings of the Workshop*. Association for Computational Linguistics (ACL), 147–156.

[52] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[53] Cecilia Lao, Jo Lane, Hanna Suominen, et al. 2022. Analyzing Suicide Risk From Linguistic Features in Social Media: Evaluation Study. *JMIR formative research* 6, 8 (2022), e35563.

[54] D. Lee, M. Kang, M. Kim, and J. Han. 2022. Detecting Suicidality with a Contextual Graph Neural Network. In *CLPsych 2022 - 8th Workshop on Computational Linguistics and Clinical Psychology, Proceedings*. Association for Computational Linguistics (ACL), 116–125.

[55] Tianshi Li, Kayla Reiman, Yuvraj Agarwal, Lorrie Faith Cranor, and Jason I Hong. 2022. Understanding challenges for developers to create accurate privacy nutrition labels. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–24.

[56] Alessandro Liberati, Douglas G Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C Gotzsche, John PA Ioannidis, Mike Clarke, Philip J Devereaux, Jos Kleijnen, and David Moher. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Annals of internal medicine* 151, 4 (2009), W–65.

[57] Huijie Lin, Jia Jia, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tang, Ling Feng, and Tat-Seng Chua. 2017. Detecting stress based on social interactions in social networks. *IEEE Transactions on Knowledge and Data Engineering* 29, 9 (2017), 1820–1833.

[58] Donald J. Maletz. 2002. Tocqueville's Tyranny of the Majority Reconsidered. *The Journal of Politics* 64, 3 (2002), 741–763.

[59] Naoki Masuda, Issei Kurahashi, and Hiroko Onari. 2013. Suicide Ideation of Individuals in Online Social Networks. *PLOS ONE* 8, 4 (2013).

[60] Naoki Masuda, Issei Kurahashi, and Hiroko Onari. 2013. Suicide ideation of individuals in online social networks. *PloS one* 8, 4 (2013), e62262.

[61] Jacob Metcalf and Kate Crawford. 2016. Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society* 3, 1 (2016), 2053951716650211.

[62] 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), 1–11.

[63] 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*. 11–20.

[64] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[65] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507.

[66] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 2053951716679679.

[67] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the Expressed Consequences of AI Research in Broader Impact Statements. *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* 1, 1 (2021), 795–806. arXiv:2105.04760

[68] Bridianne O'Dea, Stephen Wan, Philip J. Batterham, Alison L. Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions* 2, 2 (2015), 183–188.

[69] World Health Organization. 2001. The World Health Report 2001: Mental health: new understanding, new hope. (2001).

[70] World Health Organization. 2021. *Mental health atlas 2020*. World Health Organization. vii, 126 p. pages.

[71] Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2016. CLEF eRisk: Early Risk Prediction on the Internet. https://erisk.irlab.org/.

[72] 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 (2013), e217.

[73] 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 (2013), 1–15.

[74] Jessica Pater, Casey Fiesler, and Michael Zimmer. 2022. No Humans Here: Ethical Speculation on Public Data, Unintended Consequences, and the Limits of Institutional Review. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–13.

[75] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society* 7, 2 (2021), 20563051211019004.

[76] Joseph Reagle. 2022. Disguising Reddit sources and the efficacy of ethical research. *Ethics and Information Technology* 24, 3 (2022), 41.

[77] Joseph Reagle. 2022. Disguising Reddit sources and the efficacy of ethical research. *Ethics Inf. Technol.* 24, 3 (2022), 41.

[78] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ DATA SCIENCE* 6 (2017), 1–34.

[79] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6 (2017), 1–12.

[80] L. Ren, H. Lin, B. Xu, S. Zhang, L. Yang, and S. Sun. 2021. Depression detection on reddit with an emotion-based attention network: Algorithm development and validation. *JMIR Medical Informatics* 9, 7 (2021).

[81] Esteban A. Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. 2019. Anticipating Depression Based on Online Social Media Behaviour. In *Flexible Query Answering Systems*, Alfredo Cuzzocrea, Sergio Greco, Henrik Legind Larsen, Domenico Saccà, Troels Andreasen, and Henning Christiansen (Eds.). Vol. 11529. Springer International Publishing, Cham, 278–290.

[82] A. Roy, K. Nikolitch, R. McGinn, S. Jinah, W. Klement, and Z.A. Kaminsky. 2020. A machine learning approach predicts future risk to suicidal ideation from social media data. *npj Digital Medicine* 3, 1 (2020).

[83] Koustuv Saha and Munmun De Choudhury. 2017. Modeling Stress with Social Media Around Incidents of Gun Violence on College Campuses. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (2017), 92:1–92:27.

[84] Pedro Sanches, Axel Janson, Pavel Karpashevich, Camille Nadal, Chengcheng Qu, Claudia Daudén Roquet, Muhammad Umair, Charles Windlin, Gavin Doherty, Kristina Höök, et al. 2019. HCI and Affective Health: Taking stock of a decade of studies and charting future research directions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–17.

[85] R. Sawhney, S. Agarwal, A.T. Neerkaje, N. Aletras, P. Nakov, and L. Flek. 2022. Towards Suicide Ideation Detection Through Online Conversational Context. In *SIGIR 2022 - Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, Inc, 1716–1727.

[86] R. Sawhney, H. Joshi, L. Flek, and R.R. Shah. 2021. PHASE: Learning emotional phase-aware representations for suicide ideation detection on social media. In *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*. Association for Computational Linguistics (ACL), 2415–2428.

[87] R. Sawhney, H. Joshi, S. Gandhi, and R.R. Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. Association for Computational Linguistics (ACL), 7685–7697.

[88] R. Sawhney, H. Joshi, S. Gandhi, and R.R. Shah. 2021. Towards Ordinal Suicide Ideation Detection on Social Media. In *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, Inc, 22–30.

[89] R. Sawhney, H. Joshi, R.R. Shah, and L. Flek. 2021. Suicide Ideation Detection via Social and Temporal User Representations using Hyperbolic Learning. In *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*. Association for Computational Linguistics (ACL), 2176–2190.

[90] Elizabeth M Seabrook, Margaret L Kern, and Nikki S Rickard. 2016. Social networking sites, depression, and anxiety: a systematic review. *JMIR mental health* 3, 4 (2016), e5842.

[91] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution.. In *IJCAI*. 3838–3844.

[92] Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety on Reddit. In *CLPsych*. 58–65.

[93] E. Sherman, K. Harrigian, C. Aguirre, and M. Dredze. 2021. Towards Understanding the Role of Gender in Deploying Social Media-Based Mental Health Surveillance Models. In *Computational Linguistics and Clinical Psychology: Improving Access, CLPsych 2021 - Proceedings of the 7th Workshop, in conjunction with NAACL 2021*. Association for Computational Linguistics (ACL), 217–223.

[94] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, III Daumé, H., and P. Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, CLPsych 2018 at the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HTL 2018*. Association for Computational Linguistics (ACL), 25–36.

[95] Anu Shrestha and Francesca Spezzano. 2019. Detecting Depressed Users in Online Forums. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '19)*. Association for Computing Machinery, New York, NY, USA, 945–951.

[96] P.P. Sinha, D. Mahata, R. Mishra, R.R. Shah, R. Sawhney, and H. Liu. 2019. #suicidal - A multipronged approach to identify and explore suicidal ideation in twitter. In *International Conference on Information and Knowledge Management, Proceedings*. Association for Computing Machinery, 941–950.

[97] Bryan A Sisk, Jessica Mozersky, Alison L Antes, and James M DuBois. 2020. The "Ought-Is" Problem: An Implementation Science Framework for Translating Ethical Norms Into Practice. *Am. J. Bioeth.* 20, 4 (2020), 62–70.

[98] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 56–67.

[99] DJ Stillwell and M Kosinski. 2015. myPersonality Project website.

[100] Elizabeth Stowell, Mercedes C Lyson, Herman Saksono, Reneé C Wurth, Holly Jimison, Misha Pavel, and Andrea G Parker. 2018. Designing and evaluating mHealth interventions for vulnerable populations: A systematic review. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–17.

[101] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 5 (2020), 1–53.

[102] Karen L Tonso. 1996. The impact of cultural norms on women. *J. Eng. Educ.* 85, 3 (1996), 217–225.

[103] Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. 2020. Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences. *IEEE Transactions on Knowledge and Data Engineering* 32, 3 (2020), 588–601.

[104] 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 (2015), e1003904.

[105] Jessica Vitak, Nicholas Proferes, Katie Shilton, and Zahra Ashktorab. 2017. Ethics regulation in social computing research: Examining the role of institutional review boards. *Journal of Empirical Research on Human Research Ethics* 12, 5 (2017), 372–382.

[106] Jessica Vitak, Katie Shilton, and Zahra Ashktorab. 2016. Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. 941–953.

[107] Jessica Vitak, Katie Shilton, and Zahra Ashktorab. 2016. Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community. 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, 941–953.

[108] J.T. Wolohan. 2020. Estimating the effect of COVID-19 on mental health: Linguistic indicators of depression during a global pandemic. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020* (2020).

[109] A Wongkoblap, M A Vadillo, and ... 2018. A Multilevel Predictive Model for Detecting Social Network Users with Depression. In *ICHI*. ieeexplore.ieee.org.

[110] 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 (2017), e228.

[111] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, HV Jagadish, and Gerome Miklau. 2018. A nutritional label for rankings. In *Proceedings of the 2018 international conference on management of data*. 1773–1776.

[112] Amir Hossein Yazdavar, Hussein S. Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media. *Proceedings of the ... IEEE/ACM International Conference on Advances in Social Network Analysis and Mining. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining* 2017 (2017).

[113] Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. In *The Ethics of Information Technologies*. Routledge, 229–241.

[114] Michael Zimmer and Sarah Logan. 2021. Privacy concerns with using public data for suicide risk prediction algorithms: a public opinion survey of contextual appropriateness. *Journal of Information, Communication and Ethics in Society* (2021).

[115] H. Zogan, I. Razzak, S. Jameel, and G. Xu. 2021. DepressionNet: Learning Multi-modalities with User Post Summarization for Depression Detection on Social Media. In *SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, Inc, 133–142.

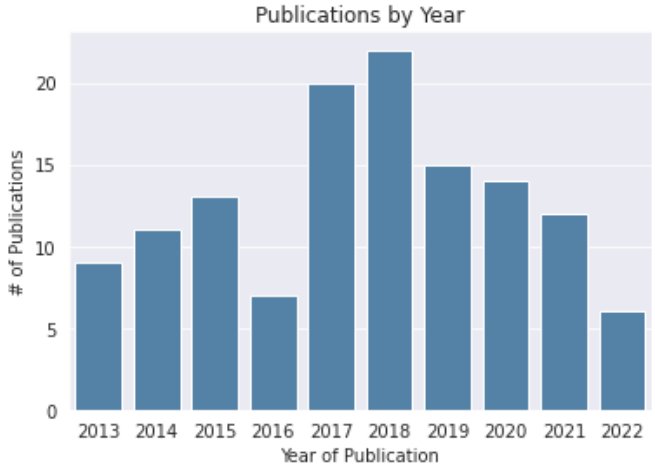


Figure 5: Distribution of publications in our SLR (n = 129)

	Coding
Human-Subjects	
Data Source (Public v. Private)	Public; Private; Inferred; Unclear
Interaction with Subjects	Yes; No; Not Described
IRB/Ethics board Mentioned	Yes, they got it; No, they didn't get it; Not described
Consent	Yes, it's mentioned; Not described; Inferred
Compensation	Dollar amount (converted to USD); Not described
Plan to inform subjects or gatekeepers	Yes; No; Not described; Not relevant
Data Centered	
Subject inclusion/exclusion criteria	Write-in criteria; Not Described
Modification of personal data	Yes; No; Unclear; No personal data included
Data de-identified	Yes; No; Not described; Unclear
Data sharing protocol	Yes, will share data; No, will not share data; Not described; Unclear
Data Storage	Yes; Not described
Impact	
Harms consideration	Yes; Not described; Not explicitly stated
Ethics disclosure section	Yes; Unclear; Not described

Table 2: Coding options for each disclosure