Disordering Datasets: Sociotechnical Misalignments in AI-Mediated Behavioral Health

VAROON MATHUR, AI Now Institute, New York City, New York, USA
CAITLIN LUSTIG, University of Washington, Seattle, Washington, USA
ELIZABETH KAZIUNAS, Indiana University Bloomington, Bloomington, Indiana, USA

The application of artificial intelligence (AI) to the behavioral health domain has led to a growing interest in the use of machine learning (ML) techniques to identify patterns in people's personal data with the goal of detecting—and even predicting—conditions such as depression, bipolar disorder, and schizophrenia. This paper investigates the data science practices and design narratives that underlie AI-mediated behavioral health through the situational analysis of three natural language processing (NLP) training datasets. Examining datasets as a sociotechnical system inextricably connected to particular social worlds, discourses, and infrastructural arrangements, we identify several misalignments between the technical project of dataset construction and benchmarking (a current focus of AI research in the behavioral health domain) and the social complexity of behavioral health. Our study contributes to a growing critical CSCW literature of AI systems by articulating the sensitizing concept of *disordering datasets* that aims to productively trouble dominant logics of AI/ML applications in behavioral health, and also support researchers and designers in reflecting on their roles and responsibilities working within this emerging and sensitive design space.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: behavioral health, mental health, training data, datasets, machine learning, artificial intelligence, healthcare, reflexivity, situational analysis, critical data studies

ACM Reference Format:
Varoon Mathur, Caitlin Lustig, and Elizabeth Kaziunas. 2022. Disordering Datasets: Sociotechnical Misalignments in AI-Mediated Behavioral Health . *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 416 (November 2022), 33 pages. https://doi.org/10.1145/3555141

1 INTRODUCTION

Artificial intelligence (AI) techniques like machine learning (ML) are widely perceived by experts across fields like computer science and medicine as having significant potential in supporting the diagnosis and treatment of people living with behavioral health conditions like depression, anxiety, schizophrenia, and bipolar disorder. Algorithms from the field of natural language processing (NLP), for example, are increasingly used in a variety of digital health interventions from conversational agents (i.e., chatbots) created to emulate interactions between a patient and a therapist [39] to clinical decision support systems being integrated into hospital electronic health record systems in order to identify and calculate the risk of mental health concerns such as suicidality and depression among patients [57, 64]. Outside of formal healthcare domains, technology companies also employ

Authors' addresses: Varoon Mathur, v3.mathur@gmail.com, AI Now Institute, New York City, New York, USA; Caitlin Lustig, University of Washington, Human Centered Design & Engineering, Seattle, Washington, USA; Elizabeth Kaziunas, Indiana University Bloomington, Department of Informatics, Bloomington, Indiana, USA, ekaziuna@iu.edu.

2573-0142/2022/11-ART416 $15.00
https://doi.org/10.1145/3555141

similar algorithms on users of their social media platforms to detect language within posts that might suggest mental health concerns, like suicidal ideation, or even integrate “clinically validated self-assessment” tools into search features to help users get additional information and resources about behavioral health symptoms [1, 62].

The rapid proliferation of AI technologies into sensitive domains like behavioral health across formal and informal health contexts has also led to growing public concern and interdisciplinary investigations into a wide range of ethical and social challenges, including people’s right to privacy, data security issues, the impact of automated systems on the patient-clinician relationships and trust in healthcare decisions [19, 71]. Machine learning research in behavioral health has often focused on questions related to bias and safety, such as how algorithms are validated before being used by or on patients. Technical approaches for addressing ethical concerns often rely upon better understanding data. ML studies of algorithmic bias, for example, typically audit machine learning datasets and large text corpora that are used for training data. Generally, large datasets are central to the development of AI systems, not only in their use for training algorithms, but in that they also serve as “benchmarks” for AI model evaluation both in academic and industry settings, defining the state of the art and validating the use of particular AI techniques for identifying and predicting different behavioral health conditions [14, 42].

While a technical analysis of datasets can be useful for elucidating the limitations of algorithms and predictive capacity across axes such as gender and race, recent critical scholarship in CSCW, STS, FAccT, and related fields have argued for the necessity of also attending to the social dimensions of machine learning data beyond technical audits, particularly in light of the growing awareness of societal harms caused by AI technologies [4, 36]. As others have noted, training datasets are not neutral information artifacts, but can be understood as parts of complex sociotechnical systems connected to particular social worlds, political regimes, histories, cultural discourses, and infrastructural arrangements [4, 32, 73, 74, 86, 87]. Understanding such data relations can help identify the (often invisible) values and assumptions embedded in the design of AI systems, and support the necessary reflection work of data scientists, designers, and researchers whose priorities and interests are also being served in the development and deployment of this technology [41, 46, 78].

Building upon this critical discourse, in this paper we aim to situate AI/ML data within their wider social context through an interpretivist study of how datasets are created and used for training and benchmarking machine learning algorithms in the behavioral health research space. Drawing on the sociological "theory/method package" of situational analysis [21], we unpack datasets as an inextricable part of the social worlds of the scientists who are building and analyzing them. Specifically, we investigate three different datasets used by a ML research community–anonymized in this paper as "the Workshop"–that represent the types of data practices commonly found in AI/ML behavioral health research.

While Workshop participants were motivated to use AI technology to better support people living with behavioral health conditions–a research goal that was often intensely personal for many people involved in this community–their research activities primarily focused around the task of discovering patterns of mental illness in personal data. They worked on this task through group activities, including an annual ML dataset benchmarking contest discussed here as the "Collective Challenge," in which ML researchers examined texts ranging from childhood school assignments to social media posts hoping to detect patterns of psychiatric disorders through analyzing people’s everyday language. Despite being rooted in good intentions, we argue in our paper that this narrow technical focus leaves many important social dimensions of behavioral health unaccounted for, and even more problematically, can be seen as pathologizing the lived experience of behavioral health.

In the following sections, we position our study within related literature on AI and behavioral health by providing an overview of the state of the art in AI/ML technologies, design approaches to behavioral health in CSCW/HCI, and critical perspectives on AI/ML technologies. Next, we outline our study methods and describe the theoretical framing grounding our reflexive approach to analyzing datasets. We turn then to our research findings, detailing Workshop activities connected to creating training datasets for benchmarking ML algorithms, as well as examining three datasets used for detecting depression, suicidality, and PTSD. Finally, in our discussion, we reflect on the possible harms for patient communities stemming from sociotechical misalignments in current data practices, and our role as researchers and designers in engaging with the dangerous promises of a technology-in-progress. We propose the sensitizing concept of *disordering datasets* as a critical provocation and reflexive approach for the CSCW community that helps make visible the problematic assumptions we might have about the usefulness of AI-mediated interventions for different health conditions and patient groups.

2 RELATED WORK

2.1 State of the Art AI/ML for Behavioral Health

Machine learning applications in the behavioral health domain largely follows supervised learning methods for predicting mental health states, including depression, PTSD and suicidality [82]. Supervised learning generally relies on the labeling practices of the data that are used to train the algorithm, so as to “predict” the labels for new data. These methods generally incorporate use of patient electronic health records, sensors that gather implicit information about the individual and their environment, and other forms of biomedical data [84]. For example, algorithms are often trained on datasets where time-series information is provided for individual patients so that datasets contain information such as date and time of hospital readmission, or an official diagnosis provided by a physician. This becomes a label representing the "ground truth" for a patient’s mental health status, and thus the target variable that machine learning algorithms are trained to predict [49].

Apart from using patient health care records or other forms of biomedical data, emerging AI/ML research in behavioral health is increasingly utilizing algorithms developed within the field of computational linguistics to identify patterns within language data that support the detection and diagnosis of different mental health conditions. These language datasets are typically derived from text corpora that are text mined and scraped from public facing web-pages or social media posts. These data are then used to train conversational agents that might mimic responses normally received by a counselor or therapist, but also to infer or predict intent or emotion behind a user’s social media posts. Such technical approaches have become standardized in collaborative research spaces between computational linguistics and health care researchers, such as the Workshop. Recognizing the value of interdisciplinary expertise, the Workshop also holds events that bring together teams of clinicians and computer scientists who want to collaboratively work towards building machine learning models on "shared" datasets, a practice found in the wider AI-behavioral health research community [45]. While such benchmarks within the general field of machine learning are common and robust, these efforts have been undertaken to replicate these practices within the space of behavioral health and AI in order to advance model building and ultimately accelerate model deployment into clinical contexts [22].

2.2 Social Impact of AI and Behavioral Health

Designing and evaluating technologies for supporting behavioral health and well-being has been an area of concern in CSCW for a number of years. Much of the CSCW and HCI research on AI for behavioral health has focused on ML methods of detecting and predicting behavior health conditions using social media data. Recently, CSCW scholarship has begun to critique the positivist views of behavioral health that motivate these interventions. Our paper adds to this critical discourse by“reading the social” back into ML datasets to make visible how behavioral health data is created (and given legitimacy) by a particular community of researchers.

In HCI/CSCW literature, social media data are often seen as good representations of people’s behavioral health because they are longitudinal data generated in situ (i.e., data collected while a person is feeling distress rather than data collected in a physician’s office [29]). Uses of behavioral health systems that use social media data include identifying the typical Facebook behavior of people who are depressed [65], predicting whether someone is at risk of depression [30], predicting whether someone will have postpartum depression based on statistical analyses of Twitter data [30] and Facebook data [31], and detecting language that is pro-eating disorder on Instagram [19]. Many of the design implications of this research are focused on developing mechanisms for platforms to make timely interventions, which have been taken up by some platforms, such as Facebook’s suicide prevention AI/ML system [26].

These early attempts to help people with behavioral health conditions were system-oriented and primarily took a technosolutionist view of behavioral health interventions. Troublingly, Sanches et al. [71] found that roughly a third of papers on HCI for affective health addressed ethical issues, and only a small portion discussed beneficence or justice. But there has recently been a critical turn towards re-evaluating these past assumptions and design framings. Researchers in CSCW, the broader HCI community, FAccT, STS, and psychiatry have identified ethical concerns with behavioral health AI systems (especially those that use scraped data from social media): risks to privacy (e.g., identification); lack of informed consent and inability to opt out; unclear standards for validity and methodological rigor; population bias in training data; lack of standardization of ethical practices; lack of clarity about when developers and practitioners should intervene if users may be at risk of harm; and risks to users when systems return false negatives or false positives (see: [18, 20, 71]). Research also points to the need for more collaboration between clinicians and computer scientists: few models are validated in a clinical setting or are designed collaboratively with people who have behavioral health conditions [71].

Importantly, a growing body of interpretivist and critical HCI/CSCW research has called attention to the social complexity of people’s lived experiences with behavioral health as it relates to system design. Feuston and Piper (2018) [37], for example, have discussed the importance of a situated understanding of how people discuss mental health and illness online, noting the harms of a “coded gaze” that classifies people in ways that may not match their individual experiences. HCI/CSCW researchers like Pendse et al. (2022) have also argued for more attention to how different communities and social worlds understand (and care for) behavioral health needs, arguing for the need to design digital health systems in ways that aren’t reductive or extractive [67]. As well, Kaziunas et al. (2019) have discussed the need for system designers to reckon with complex social worlds and systemic inequities in behavioral health. They write that since “design interventions are always partial and incomplete,” designers need to recognize the way new technological systems are connected to the wider “infrastructural brokenness” of healthcare, such as stigma and the difficulties people experience when trying to access local behavioral healthcare services [53]. We build upon this scholarship by offering an empirical study of ML behavioral health datasets, and by theoretically broadening the ways CSCW researchers and designers contend with the lived experience of both patients and researchers within the AI-health design space.

3 STUDY DESIGN AND METHODS

This study is part of a larger, multi-year research project investigating the social impacts of AI technologies in behavioral health. Our research goal here was to better understand the wider social context of training data, including the activities and motivations of dataset creators. We identified a particular ML research community–which we refer to as "the Workshop"–as a site of analytic interest given its public and active role in shaping NLP research on behavioral health, both in its annual dataset challenges, but also with regards to its participants' presence at and contributions to prominent ML academic conferences.

In our data collection and analysis we followed situational analysis, a methodological approach articulated by sociologist, Adele Clarke and colleagues, which is an interpretivist extension of grounded theory [21]. Drawing on theory from Science and Technology Studies (STS) and symbolical interactionism, situational analysis or "SA" examines forms of action and the relationalities between sets of human/nonhuman actors, social worlds, and the various historical, symbolic, political, and discursive elements from an ecological perspective. Centering technofeminist concerns, situational analysis also draws explicit attention to the politics of artifacts, power of discourses, and reflexivity in the analytic research process. In this, situation analysis can be best understood as a "theory/method package," [21] that involves an iterative and ongoing process of data collection and analysis that mutually inform one another, as is common in interpretivist research. Through the course of our study, then, we drew upon a set of theoretical literature from across critical data studies and STS–which informed our critical analytic lens on the political categories of chronic illness and social meanings of data. Disability studies literature [59] also helped to focus our critical lens by presenting discursive strategies for reframing and pushing back against limited and harmful categorizations of behavioral health conditions.

As our study aims to highlight common types of data practices, we have chosen to anonymize the research community and datasets we analyze using pseudonyms. As a "critical" analysis, our paper is not intended as a critique of any one particular dataset or ML researcher, but a means of drawing attention to (and questioning) the dominant logics of AI systems and popular approaches to data science in this domain as “critical friends” [28]. We also have anonyomized any shared excerpts from the datasets themselves.

3.1 Data Collection

Through the course of our study, we collected a wide range of related documentation about the ML research space of behavioral health. This included popular media articles and reports from national funding bodies in the United States, such as the National Institute of Health (NIH), on the topic of AI/ML and behavioral health in order to identify popular narratives, scientific visions, and values. We also collected documentation, both online communications and information artifacts, specifically about Workshop activities and its wider scholarly community. This included gathering together published descriptions of dataset benchmarking activities from active and archived websites, published research papers that described methods and considerations around dataset construction and use, as well as a number of blog posts and online interviews where prominent Workshop participants discussed their research work and reflected on their personal motivations, future hopes, and concerns with using AI technologies in behavioral healthcare. Public recordings of podcasts and video talks were downloaded and transcribed.

3.2 Machine Learning Dataset Collection

We also downloaded dataset files used by Workshop participants in an annual dataset research event that we refer to in this paper as the "Collective Challenge." We chose three Collective Challenge datasets that were 1) publicly available to the wider academic community; 2) that have been used and cited by other researchers developing machine learning algorithms for behavioral health; and 3) that have specifically been used as benchmarking tasks for comparing model accuracy. Obtaining access to each dataset involved an application process that required our team to create formal submissions of our research intent and evidence of our academic credentials. Accessing permission for two of the datasets involved ongoing direct communications with the dataset authors to explain the details of our research project. We were also required to submit data use and confidentiality agreements, as well as provide documentation of Institutional Review Board (IRB) approval for our study. The other dataset creator required the creation of an online account with an institutional email in order to gain access. In our study, we have complied with all the dataset agreements, as well as obtained the required IRB approval.

Datasets	Corpus Source	Corpus Size	Prediction Task	IRB Requirement
The 'School Essay' Dataset	Digitally reconstructed from survey and questionnaire responses from school children in the 1950s, up until the age of 55	10,000+ childhood essays, 4,000+ corresponding essays at age 50	Predict current and future psychological health from an essay authored by children	Create an account on a centralized data archive service, and agree to data use and confidentiality agreement
The 'Reddit' Dataset	Data scraped from specific sub-communities within the social media site Reddit	10,000 users, 1,500,000 total Reddit posts	Predict the degree of suicide risk from a Reddit post	Contact dataset authors with IRB letter and signed data use and confidentiality agreement
The 'Twitter' Dataset	Data scraped from the social media site Twitter	600 users, 3000 Tweets per user	Predict the correct diagnosis of PTSD or depression from a Tweet	Contact dataset authors with IRB letter and signed data use and confidentiality agreement

Table 1. Summary of datasets collected from the Workshop. Each dataset consisted of a unique corpus and prediction task, and were each acquired separately.

3.3 Data Analysis

Data analysis occurred over the course of a year. In the first stage of analysis, the second and third authors used situational analysis to analyze ML literature in the behavioral health domain, as well as media articles on AI technologies for behavioral health, with the research goal of identifying different sets of relations in the AI-mediated behavioral health design space. This analysis resulted in thematic coding on popular design narratives, types of personal data being used in ML research, and known social and ethical challenges. Utilizing Clarke's situational analysis mapping techniques, we also created an array of situational, discourse, and social worlds/arena maps and analytical memos [21]. These map artifacts and memos were discussed among the entire research group as theoretical insights emerged, and informed our subsequent analysis of the Workshop community and ML datasets. In the second stage of analysis, led by the first author, we focused our attention on the Workshop and related dataset activities. Each of the three authors independently analyzed the three ML datasets (this analysis process is detailed below in Section 3.4 as a "close reading" activity), as well as the related documentation (e.g. papers, podcasts) using open-coding methods to identify significant themes. Dataset files, transcripts, and related information artifacts were then discussed collectively among the research team during data analysis sessions. New codes were generated as important concepts were identified, compared, and revised.

3.4 "Close Reading" of Training Datasets

While situational analysis can be used to analyze any document—including information artifacts like datasets—it is not a method which has been widely adopted for examining the granular details of computational technologies at the level of data structures. We therefore found it helpful to also draw on techniques found in HCI/CSCW and critical data studies for investigate the social and material configurations of software, algorithms and datasets. We incorporated several of these techniques into our larger situational analysis mapping activities [33]. In particular, we drew inspiration from Denton et al.'s [32] paper which discusses "reading the dataset as text" to investigate the “unspoken conventions” of their construction. Denton and colleagues argue that this includes examining the documentation of the dataset, and how it has been used in the academic and industry contexts [32]. Importantly, Denton et al.'s work highlights several important social dimensions around AI training data such as the motivations behind the creation of a particular dataset, and the “embedded” social norms that structure the process of data collection and curation [32]. Also informing our study, Geiger et al. [43] discussed various social and technical dimensions in the creation of machine learning datasets. Specifically examining annotation work, they look at whether or not a dataset was designed specifically for an original classification task, how people were involved in the creation of data labels, and if these human labelling processes were made transparent. Their work argues for viewing human activities as a crucial part of how such datasets help establish the scientific validity for models that are trained on top of them [43].

Drawing inspiration from these studies, we conducted a “close reading” of each dataset as a means of better understanding the ways people's everyday activities are collected, classified, and documented as ML behavioral health training data. In this analysis, we paid close attention to the engineering values and assumptions around behavioral health conditions that influenced the design of these datasets. We began our close reading sessions by noting various technical elements of each dataset from a data science perspective. This included examining the types of data being captured and how datasets were formatted to better serve model development (e.g., how much of the dataset would have needed to be cleaned of incorrectly formatted data). Next, we discussed the social significance of each data element and their connection to particular forms of disciplinary training, organizational contexts, shared social norms, and histories. For each of the three datasets, we also analyzed a wide range of related documentation that provided insight into how each dataset was created, understood by members of its social world, and shared, including: README files, websites and blogs, IRB applications, and associated academic research papers that used the datasets. This documentation helped make visible the shared conventions and routine processes around machine learning dataset creation and use in behavioral health ML research, such as the application steps needed for approval in sharing datasets with sensitive health information, the maintenance of open source repositories, as well as how datasets can be used as potential benchmarks to inform the future work of the field.

Through our close reading sessions we actively questioned the relationship between social and technical dimensions of ML datasets, asking: How were user metadata being treated? What was the value of a dataset for different members of the Workshop community—e.g., was it seen as a clinical artifact, or an exercise in building computational linguistic models? What was the desired impact of the dataset construction and dissemination for those creating it? Such questions helped focus our analysis and led us to examine which (if any) psychometrics informed the creation of each dataset, as well how widely Workshop datasets were being used as benchmarks in the wider research community. After completing this process of close reading excerpts of each training dataset and its related documentation, we then incorporated our findings of specific data practices and their related values/assumptions into a more typical SA map of the broader "situation" of AI-mediated behavioral health.

3.5 Reflexivity in Interpretivist Data Science Research

Researchers have recently called for data science research to embrace reflexivity [16, 80]. Reflexivity can generally be thought of as a process by which researchers analyze how their identities and experiences impact their research. Finlay [38] identifies multiple benefits to reflexivity as a tool that can help researchers to:

- “Examine the impact of the position, perspective and presence of the researcher.
- Promote rich insight through examining personal responses and interpersonal dynamics. Open up unconscious motivations and implicit biases in the researcher’s approach.
- Empower others by opening up a more radical consciousness. Evaluate the research process, method and outcomes.
- Enable public scrutiny of the integrity of the research through offering a methodological log of research decisions” [38]

We found situational analysis [21] to be one starting point for “doing” reflexivity beyond positionality that helped our team articulate a number of commitments and vulnerabilities, both with one other as researchers, but also in terms of our relationship to the project of AI-mediated behavioral health. Clarke et al. [21], for instance, asks researchers to “stay with reflexivity” throughout the analysis process in order to understand how they fit into *the situation* or site of analytic inquiry. This involves reflecting on how a researcher is embedded within power relationships, in deciding which actors and actants are implicated or silent, and in how a researcher’s perspective influence what is analyzed. Clarke et al., write: "When you include yourself(ves) on the initial situational map, when you analyze your relations with various other elements on that map, when you are surprised, upset, or gleeful during your research, all of these reactions matter. Understanding how and why and what the implications are for your research project is part of the process, not some external "noise" or bias" [21].

In our study, we adapted the following questions from Clarke et al. [21] to guide our reflexive process. On researcher visibility, we asked: *Who is the researcher? How is the researcher positioned vis-a-vis the situation of AI-behavioral health? Whose knowledge about AI/ML, data, or illness “counts” to whom, and under what conditions? Who/what is being researched in ML behavioral health datasets? Why and with what consequences? Who/what may be placed at risk by this research (both AI-behavioral health dataset creation, as well as our own critical study)?* Additionally, drawing on Clarke et al., we examined dimensions of difference through such questions as: *To what extent do researchers present different perspectives on AI-behavioral health in their study–even perspectives we disagree with? Who/what is omitted or silenced by researchers themselves? How are contradictory responses addressed?* These structured questions prompted us to reflect and articulate meaningful aspects of our lived experiences in relation to studying the "situation" of AI-mediated behavioral health and activities like ML dataset creation. As we later detail in the paper’s discussion, a reflexive approach was analytically useful for grappling with our multiple roles as CSCW researchers, but also as patients, caregivers, and concerned citizens; and for highlighting the different responsibilities, vulnerabilities, and stakes those roles brought with them in making visible and addressing the various sociotechnical misalignments of AI-mediated behavioral health.

3.6 Limitations

This paper reports on the documents and data activities of one ML research group and three related training behavioral health datasets created by this particular community. In part, this focus was a reflection of our research questions, as we were interested in exploring the social context of dataset creation. The number of datasets we analyzed, however, also reflects pragmatic challenges in obtaining research access to a wide range of ML training data. Datasets used in industry, for example, are often proprietary, while clinical datasets derived from the U.S. context are HIPAA protected. While we were limited in the scope of this current project to studying datasets that were open and shareable, the ML training dataset examples we detail here reflect technical approaches to detecting behavioral health conditions in personal data widely adopted across the broader research field. Moreover, as an interpretivist study, our primary goal is to offer the CSCW community theoretical insight and a critical lens for examining ML datasets, rather than provide a systematic or exhaustive review of all training data in this domain.

4 STUDY FINDINGS

In this section, we first situate machine learning (ML) datasets within their social worlds, examining both the research activities around creating training data, as well as the underlying motivations and data science logics at work in the application of AI technologies to behavioral health. Next, we turn to detailing three text-based datasets which are representative of the types of data collected and AI/ML techniques (e.g., Natural Language Processing or NLP) being used to research behavioral health. While such datasets are often created with well-intentioned goals for helping develop AI-driven systems that can one day help address pressing challenges–such as a need for greater patient support and faster access to healthcare services–they can also embed problematic biases regarding the representation of mental health conditions, as well legitimize narrow technical approaches to understanding socially complex health concerns. Grounding our analysis with empirical examples from each dataset, we highlight several important misalignments between the technical project of creating ML datasets, the social dimensions of behavioral health as a lived experience, and the long-held CSCW ambition of using technology to improve and support patient needs.

4.1 The Social Worlds of AI/ML Behavioral Health Research

There has been widespread interest in using AI technologies to help address long-standing behavioral healthcare gaps and patient needs with digital health interventions such as real-time clinical monitoring of patients to automated tools supporting people's management of common conditions like depression, anxiety, and bipolar disorder across a number of different research communities. Academics from fields like computer science and computational linguistics, engineers and industry experts from small and large tech companies, clinicians and healthcare providers, as well as national funding organizations like the National Institute of Health (NIH) in the United States have all helped create an active research space around AI-mediated behavioral health. Recognizing the complexity of this domain, a number of AI-behavioral health conference events and collaborative projects have been created over the last decade in an attempt to bridge these diverse research groups and share expertise across different fields.

One such popular AI-behavioral health event (that we have focused on in our study) is "the Workshop." Held annually at a machine learning conference venue, the Workshop aims to bring data scientists, engineers, and computational linguists together with clinical experts (which in this case is primarily psychiatrists and psychologists). Since its inaugural year, the Workshop has also held an annual dataset competition known as the "Collective Challenge." This event involves teams of researchers from both industry and academia who use various ML techniques on a shared dataset in order to see which model performs best. Each year, the Collective Challenge organizers–who are often computer science PhD students or professors working in this research space–identify a particular behavioral health condition and provide the group with a new corresponding dataset. Past events have analyzed text-based datasets with the goal of trying to accurately detect and predict conditions like depression, bipolar disorder, anxiety, and schizophrenia. The text-based datasets, as we will discuss in more detail later, range from childhood school essays to social media data scraped from popular public sites like Twitter and Reddit. Social media data in particular holds a special promise for Workshop participants (and for the wider research community studying AI and behavioral health) as it provides valuable observational data about the crucial time people spend in between clinical encounters–what some researchers in this field have called a "clinical whitespace" [88]. Analyzing this data, so the design narrative goes, can provide researchers with insight needed to create tailored AI-driven digital health interventions for improving patient care.

The corpus of text determines the type of behavioral health intervention and the specific technical task identified as the challenge for that year. For example, as we will detail later in Section 4.2, a Twitter dataset might focus on predictive tasks for depression and PTSD detection because they are understood by researchers as conditions found more commonly on social media platforms compared with other behavioral health conditions. A Reddit dataset, scraped from the subreddit community, r/SuicideWatch, would have a predictive task focused on suicidality given the assumption that users who post here are self-disclosing, to some level, personal experience with suicidal ideation.

While the Workshop is organized like a typical professional academic conference event, in many ways the Collective Challenge also shares a similar ethos to popular digital health events like healthcare hackathons in which multiple teams compete to address a healthcare "problem" through applying their coding skills and engineering expertise. And while many popular digital health coding events, such as MIT's Hacking Medicine, have sought to include the perspectives of patient advocates, we did not see evidence of patient participation among the social worlds of AI-behavioral health research. Though the Workshop was intended to be an interdisciplinary research space–bringing diverse behavioral health experts together–at the time of our study, we found the Collective Challenge was primarily composed of teams made up of computer and data scientists. As we discuss next, the dominant and privileged position of data science in the social world of the Workshop is significant in how the "challenge" of behavioral health becomes framed almost exclusively as a technical issue or "information signal problem."

4.1.1 Dataset Logics: Framing Behavioral Health as an Information Signal Problem. In the field of AI-behavioral health research, chronic illnesses like bipolar disorder and anxiety are often approached as technical puzzles that could be solved with access to enough personal data, collective tinkering, and a very accurate model. A central claim underlying the dataset activities of AI-behavioral health researchers (and canonized in ritual events like the Collective Challenge) is that people living with conditions like bipolar disorder reveal information about their health conditions through their conversations and social interactions, and that analyzing different types of digital communication (like social media posts) has the potential to reveal these specific linguistic patterns. For researchers participating in the Workshop, then, there is an underlying belief that the complexity of mental illness can (to some degree) be reduced to a technical problem of identifying the right kind of information "signals."

The creation of ML behavioral health datasets are influenced by several disciplinary logics, including information science, medicine, and linguistics. The task of extraction and detection of linguistic "signal" from contextual "noise," for example, comes from classical signal processing techniques in engineering subfields [11]. Specifically, the framing of detection of mental health signals in language, and how human language is considered a "discrete/symbolic/categorical signaling system" that is found widely in NLP literature [83] stems from computational approaches to information transfer.

The usefulness of these datasets, then, is predicated by a view that people living with behavioral health conditions exhibit different linguistic patterns than a general population when communicating verbally and through writing. Similarly to how biomarkers are conceptualized in traditional biomedical spaces as indicators for disease onset, particular linguistic patterns or signals–such as the number of topic changes or pronoun usage–can indicate mental health issues or growing behavioral health concerns. The primary motivation of the creation of Collective Challenge datasets, then, is to validate the idea of AI/ML technologies being able to identify meaningful signals related to behavioral health from various types of personal data, including social media posts.

Another disciplinary logic underlying the use of ML text-based datasets comes from the field of psychiatric medicine, which views patient language as foundational in the diagnostic process. In our study, we found dataset creators commonly adopted and used data from clinical instruments (e.g., PHQ-9 for detecting depression) or neuropsychological assessments (e.g., the Controlled Oral Word Association test (COWAT) for evaluating schizophrenia) that were originally designed for diagnosing and monitoring patients. While the clinical instruments and assessments are generally accepted by the medical community as a reliable method for detecting behavioral health conditions across different patient populations, they also have particular limitations in guiding technology design. For instance, the usefulness of diagnostic instruments in clinical practice is highly situated and test results are impacted by several social factors, including the experience and skill of the healthcare professional trained to interpret the nuances of patient language, as well as the trust patients have with clinicians in answering structured set of questions about sensitive health experiences. An impetus for driving much of the ML dataset work in behavioral health (and collaborative efforts between computational linguists and clinician researchers more generally) has been the opportunity to automate these traditional clinical assessments in light of the chronic shortages of trained clinicians in many behavioral healthcare systems.

4.1.2 Dataset Practices for Determining Dataset Validity. Interestingly, "signal" detection in language data is not a common or traditional methodology within the field of clinical psychology or psychiatric medicine. In fact, the technical practice of signal processing is entirely separate and distinct from signal detection "theory," which is an established practice in psychology used to understand individual decision-making behavior. Given differences in disciplinary traditions informing ML datasets, we turn next to describing several common activities for ensuring the machine learning models developed and trained using the datasets will be valid in the field of clinical psychology.

Models trained from these datasets aim to make accurate determinations about an individual's behavioral health status. Importantly, for Workshop participants, and others in the fields of NLP and computational linguistics, there is a distinction made between "explicit" and "implicit" signals. Explicit signals encompass such language that is used to construct the training set in the first place (e.g., "I have been diagnosed with X..."), whereas implicit signals are defined by patterns in the language that appear consistently over time and are correlated with behavioral health outcomes. For example, one group who participated in the Collective Challenge wrote about their work developing NLP models trained on a Twitter dataset that identifies word clusters associated with users who are known to be dealing with depression, such as a higher frequency of words that can be described as seeking "personal attention" or the higher frequency of swear words in tweets. While such computational techniques for signal prediction and extraction present an interesting technical problem for ML researchers, of equal concern for Collective Challenge participants given the sensitivity of domains like behavioral health is if these models are also clinically accurate, and how scientific validity across different fields can be determined.

Given these legitimate concerns, an important documentation practice in each Collective Challenge is the creation of dataset notes that aim to provide teams with a range of contextual information to consider about the behavioral health condition or particularities of the dataset before they begin their analysis. For example, dataset notes might provide details on why populations are segmented by age and gender, an important factor given that mental health conditions themselves vary between demographics. Dataset notes might also describe the usefulness of demographic classification tools (a method to algorithmically determine the age and gender of someone posting on social media) given that such controls are common practice in clinical psychology. Other important dataset information includes control variables, such as gender and social class to account for "confounding" effects. Additionally, dataset notes often cite the need for any next steps to include randomized control trials to clinically validate models trained using the dataset itself. As well, some dataset notes discuss methods to include more clinical validity by utilizing dataset annotations by clinical experts as well as crowd sourced annotations in order to compare performance between different types of labeling. The careful attention given to creating dataset notes in the Collective Challenge demonstrates how social worlds like the Workshop seek to make traditional engineering practices relevant and meaningful to clinical psychology. As we will explore, however, the dominant framing of behavioral health as a computational challenge of signal detection and the focused concerns around clinical validity also leaves important social factors out of scope in AI-behavioral health research. One critical absence is that of people's lived experience of behavioral health and its significance to the project of dataset construction.

4.2 Sociotechnical Misalignments in AI-Mediated Behavioral Health

We now turn to analytically unpacking three ML datasets used in the Collective Challenge that are also typical of other datasets used in this field. Rather than offering a systematic or comparative review, the interpretivist study findings we share here (summarized in Table 2) aim to generate questions and reflections about the relationship between the narrow technical project of AI-mediated behavioral health and the wider social dimensions of behavioral health as tied to particular historical, environmental, and sociocultural contexts. In doing so, we make visible the different types of "sociotechnical misalignments" that can arise through current practices of dataset construction and benchmarking, and which are relevant in how we conceptualize the AI-health design space in CSCW.

4.2.1 The School Essay Dataset. The first Workshop dataset we examined was constructed from digitized text data originally collected as part of a longitudinal study conducted by the UK government. In 1958, the National Child Development (NCD) Study collected essays and questionnaire responses from a large cohort of British schoolchildren [68] and followed them over the course of a 60-year period. The original purpose of the NCD Study was to investigate various clinical and social factors associated with early childhood mortality, but the study grew in scope over the years to also investigate health problems occurring later in life. This longitudinal study comprises an enticing corpus of text data in the form of personal essays from different time points in people's lives, combined with psychometric scores taken at those time points to provide measures of the children's psychological health over time [56].

The digitization of these written essays and questionnaire responses provides a unique dataset and benchmark for machine learning models, particularly because datasets of this size are rarely accompanied with "high-quality" psychometric assessments at multiple timepoints for each individual. In order to build algorithms to accomplish predictive tasks, "labels" within datasets are

Datasets Examined	Misalignments in Construction	Misalignments in Application
The 'School Essay' Dataset	- Cultural and temporal contexts are seemingly flattened - Potential for "poverty" as a proxy for being troubled given socioeconomic classifiers used as numerical variable - Little transparency to understanding the exclusion/inclusion criteria for the Benchmark, given that it's a digital reconstruction of another dataset	- Models use language as "signals", but this language from the 1950's use is out of step with how children communicate now - Little understanding with how algorithmic interventions for children might shape their journey within the present health care system - Given the benchmarks' use of poverty to assess wellbeing, there are serious concerns with the papers' conclusions that such a model can be deployed further in resource-poor settings for "clinical use"
The 'Reddit' Dataset	- Crawling internet subcommunities that are designed for people to discuss and disclose mental health concerns biases models - Suicide degree rating from data labelers are potentially biased given that discernment of "thoughts" and "feelings" is necessary to make such assessments - Reddit is skewed demographically, which is not a consideration baked into model deployment	- Surveillance technology empowers clinicians and health system monitoring over the needs of patients, and does not account for how these tools then impact how users behave online with this knowledge - Models are trained and benchmarked to identify at-risk patients, but the dataset contains posts from users who describe how they have identified themselves to be at risk and have not received the care they need, reinforcing a cycle in a broken healthcare system - Labeling of "degrees" is motivated in part by lack of capacity for health systems, and not from a more objective understanding of severity
The 'Twitter' Dataset	- Automated labeling of gender and age creates false labels taken as "ground truth" for prediction tasks - Dataset is constructed with a bias towards "complete" data - Benchmark does not account for inclusion of bot activity	- Surveillance technology empowers clinicians and health system monitoring over the needs of patients, and does not account for how these tools then impact how users behave online with this knowledge - Models use words and phrases to identify "signals", but they do not account for how subcommunities use and communicate on social media differently - Models trained on self-disclosure inherently disconnected from identifying depression or PTSD from those who would not disclose or have different activity levels on social media in the first place

Table 2. Summary of misalignments in dataset construction and intended applications for subsequent machine learning models uncovered from close readings.

used to help guide the process of identifying patterns within the data. Such labels are intended to create a "ground truth" that models can then be benchmarked and measured against. For this dataset, training labels come directly from the original NCD study, which uses several quantitative scales for measuring behavioral health: 1) a score derived from a standardized psychometric test, the Bristol Social Adjustment Guide (BSAG), to measure psychological health at age 11; along with 2) a score derived from another psychometric scale, the Malaise Inventory, as a measure of psychological distress at ages 23, 33, 42, and 50. This type of labeling differs markedly from the datasets constructed from social media data, as we will explore later, in which such metrics are not available. In this paper, we refer to the repurposed NCD Study text and labels for the Collective Challenge as the "School Essay" dataset.

Though the dataset was designed as a Collective Challenge for the Workshop, its impact has rippled beyond the academic context. Findings from different team submissions have been used to share the positive impact of the original NCD study to those that contributed questionnaire responses as children, as well as communicating the use of “AI” as a positive incentive for the UK government to continue to fund such research through the Center for Longitudinal Studies [3].

Prompt: Imagine you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to write)	**Psychometric Score from BAGS**	**Subscale score for Anxiety, derived from BAGS**	**Subscale score for Depression, derived from BAGS**	**Gender**	**Label for Father’s Occupation**
When I am twenty five I would like to be a droughts man because I like drawing. I would like to live in a house near the river so I could go fishing a lot. I would like to go on cabin cruesers to far off countrys then when I leav the droughts mans job I would like to join the navy because I like P.E. and the ships it would be exciting every time I go on a cruese I would like to be on target practice and shoot eny old ships and sink them.	4	0	0	0	III Manual (A skilled manual occupation)
I like to look after animals, in my spare time I like sewing, I like to read books and whatch telivition. I work in a hospitall and hope to be matron one day. I like to buy nice dresses and buy nice shoes as well. I live in a flat a very nice one too, i like to read cowboy books and whatch cowboy filmes. I like to ride my bicicly. i like to draw* pictures and hang them up on the wall. I like being a nurs, i have anley down 2 cososs. I *keep all sorets of fish, like tadpolls, goldfish, rainbow fish. I have brown hair and blue eyes I like read letters from people.	57	8	6	1	III Manual (A skilled manual occupation)

Table 3. Excerpt from the “School Essay” dataset, derived from the UK’s National Child Development Study. Columns right to left show excerpts of the children’s essays, three psychometric scores given to the children by their teachers in the 1960’s, and classifications for gender and father’s occupation.

AI and NLP researchers participating in the Workshop view everyday communication and writing, like these childhood school essays, as holding a wealth of valuable information on people’s various states of psychological well-being, including anxiety and depression. The Collective Challenge task associated with this dataset asked teams to explore how analyzing a person's early language with ML techniques could be used to find "signals" for future behavioral health outcomes. A close reading of the School Essay dataset, however, highlights the importance of grappling with the wider sociohistorical context of ML training data, especially given the culturally specific and temporally-bound ways behavioral health conditions are both experienced by study participants and formally classified.

For example, excerpts from the School Essay dataset (see Table 3) demonstrate how the lived experiences of two different British school children from the 1960's are associated with specific psychometric evaluations. As part of the original study, UK researchers gave children an open-ended prompt to reflect on: "*Imagine you are now 25 years old. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 minutes to write).*" Scrolling through the dataset Excel file that organizes these essays in neat columns, children write about wanting to join the Navy, watch cowboy films and television, read letters, and buy nice shoes. "*I would like to live in a house near the river so I could go fishing a lot,*" writes one student imagining a grownup life. Another child pens, "*I like to ride my bicicly [bicycle]. i like to draw pictures and hang them up on the wall.*" Knowing that these school essays come from the UK in the late 1960's can help explain why obscure job titles like "a droughts man" and "matron" are mentioned by children, but more importantly, it also points to the ways genres of communication–like a school essay–are complex information artifacts reflecting a particular time, class structure, and cultural geography.

Re-situating this dataset back within its sociohistorical context prompts one to consider what types of information get taken up (and left out) in the process of transforming school essays into datasets that train NLP models to predict complex health states like depression and anxiety in present-day populations. For ML language models trained on historic data, what does "validity" mean when computational techniques span not only different scientific disciplines, but seek to flatten cultural and temporal distances? In a sensitive domain like behavioral health, the risks of a problematic proxy or of spurious correlation being incorporated into the dataset have serious medical implications, as well as heavy social costs. Such risks are further intensified by using non-contextualized historic behavioral health data.

Take, for instance, how the essays in the School Study dataset are connected to the child's corresponding psychometric profile, as determined by a BSAG test. Designed to "*obtain a picture of the child's behavior in a school setting,*" the total score reported, as well as the subscale scored on Anxiety and Depression included in the Workshop dataset, can be interpreted as "*higher the score, the more indications there are of problem behavior*" [75]. Table 3 shows an example of two different BSAG profiles representing students with differing genders according to a binary classifier, but with equal socioeconomic designations based on each child's father's occupation. At first, the essays themselves–two childhood fantasies of playing with pets and going off to sea–don't seem to hold any explanatory power for illuminating the widely divergent scores of a 4 and 57. Researching the various factors that go into a determining a BSAG profile, we can learn that the occupation of the student's father is used as a proxy for social status, but this fails to consider other potential key factors such as a mother's salary, family's self-reported ethnicity, and size of household that also play a crucial role in estimating socioeconomic outcomes. As well, though BSAG scores are described as being validated by more than one teacher at the child's school, no further information is provided as to how social and cultural bias might play a role in such scoring of students [75].

In light of the complexity of such social factors, it is easy to see how being from an economically disadvantaged family might problematically become a proxy for "a troubled child" in the classed, raced, and gendered classroom dynamics of 1960's Britain, especially given the lack of specialized training among the primary school teachers. This complicated social context, however, is not visible in the Workshop's use of ML datasets like "School Essays" where the hunt to find predictive relationships between a young student's school assignment and an adult's present-day behavioral health needs–no matter how fraught–is prioritized over potential future harms.

These misalignments in the datasets construction propagate further into its imagined use cases. The dataset authors point towards steps that can move these models into "clinical use" by highlighting that how the utility of machine learning models trained on childhood essays using similar psychometric assessments might be "most valuable where administering detailed assessments is particularly costly or burdensome" [56]. Though there are concerns how this particular dataset may be disproportionately biased against children from immigrant families or lower socioeconomic standing, models are still contextualized as potential solutions for resource-poor settings.

4.2.2 The Reddit Dataset. The second Workshop dataset we examined was created from text scraped from a dedicated online forum on the social media site Reddit. Reddit has become a popular source of free public data in the ML research community for those building different types of behavioral health training datasets. Within Reddit, different "subreddits" exist as communities or hubs for users interacting within the site. This particular subreddit community focuses on supporting group discussions around personal experience with suicidal ideation. The dataset includes users who posted extensively across a number subreddit communities, in addition to making posts in r/SuicideWatch.

The use of machine learning to address and support suicide prevention is often cited in the broader computer science research field, industry, and popular media as an example of using AI for "social good" [51, 72]. The Workshop's stated goals for creating a subfield around NLP and psychiatry unquestioningly embraces this design narrative, seeing the Collective Challenge datasets as a tool for pushing the state of the art forward and well as a means of shaping the future of clinical care for the better. In the Workshop's social world, as with the broader AI-health research field, computational techniques are a much needed solution to the growing political and economic burdens of untreated mental health conditions. This motivation is especially strong among researchers using AI technologies that aim to detect suicidality in online communications with the goal of protecting lives.

The Reddit dataset was created to specifically focus on developing more robust machine learning models that are able to predict the "degree" of risk for suicidality. The dataset creators write, for instance, that they were motivated in part by the fact that the constraints of our current behavioral healthcare system (e.g., lack of resources and trained specialists) requires the prioritization of people who are most at risk, and binary classification tools do not provide such specificity. These risk labels are annotated by experts, who use a standardized rubric to assign a degree of risk based on several factors, such as thoughts or feelings expressed, and whether or not methods of attempting suicide are outlined within a user's post on r/SuicideWatch. For each user included in the dataset, the collective posts from other subreddit communities they may have posted in are also included in order to add context for individual users represented in the dataset.

Our close reading of the Reddit dataset highlighted how the activies of dataset construction were often in misalignment with the ML researchers' stated goal of building early warning systems for suicidality. While the psychometric profiling of Reddit users based on degree of risk (once again focusing solely on the use of language as a signal) raises important questions around scientific validity and *what* is being measured; our focus here is on what these data practices reveal about what communities like the Workshop understand as the "usefulness" of computational techniques as interventions for helping fix a broken healthcare system, and what types of care activities should be prioritized and optimized. In other words, what clinical utility models are envisioned when trained on Reddit posts?

subreddit	post_title	post_body
depression	thinking about going to hospital	I've been extremely depressed for years and feeling suicidal since school started. A few days ago I made an appointment to talk to my school counselor about it. He mentioned that if I felt too overwhelmed with the suicidal thoughts, I would go into one of those mental hospitals or inpatient clinics (I don't even think I'm using the right name). And now I'm feeling really overwhelmed again with dark thoughts...so should I go? My friend went once. And from what she told me it sounds like it might help. But the thing I'm freaked out about is if I am locked up and can't leave if I want. All my friends are online and if I'm scared I wouldn't be able to go online for a long time. And that would be horrible. But I don't know what else I can do. It seems like a better option than suicide.
SuicideWatch	lying for weeks now.	I've been depressed ever since I was about 12 and on and off suicidal since I was 13. Recently, ever since my 30th birthday, I can't stop thinking of just ending it all. I've lied to my therapist and family about it. My mom always guilted me into not doing it in the past, and my dad doesn't say anything. I haven't told my boyfriend either. I just want to remember what it's like to feel happy l for once in my life.
confession	Need to talk to someone	I've been taking Paxil for my depression, even though I really didn't want to go on another prescription medication. I hate the side effects, but going off it even for a few days is so painful. I feel like crying all the time and feel so depressed because nothing works. I can't get any pot and it's the only thing that helps me. Sorry. I needed to vent... If things stay this way, I'm not sure I can keep going.
Depression	In so much pain.	I feel so angry. Depression consumes every good part of my life. My girlfriend just dumped me. My career is falling apart. The last three years have felt like one giant downward spiral that I can't stop. I want a therapist, but I have no money. I put a gun to my head tonight. I just want it to stop.
genderqueer	asked my doctor to refer me to a gender specialist and got no help.	After a really pathetic suicide attempt I went to my doctor for help. He put me on antidepressants again and got me into crisis counselling for gender issues. I'm confused about what gender I am and felt a part of me was being ignored.My therapist told me that I should speak to a gender specialist who can help me figure this out, so I asked my doctor for help. His referral came and it was for. . . . a sexual dysfunction clinic.I am deeply upset and feel triggered by this. Or am I just being overly sensitive?

Table 4. Excerpt from the Reddit dataset, edited for anonymization. Columns left to right show subreddit, post title, and edited excerpts of user posts.

The lack of access to quality behavioral health screening, and the perceived lack of clinicians to properly assess and triage high-risk individuals represent particular political and economic narratives for the development of these datasets. These narratives seem to focus on "scale," and on the promise of augmenting physicians' ability to monitor the patient populations they serve. However, we find that these assumptions don't necessarily align with the experiences of people living with behavioral health conditions who make up these datasets. Compare, for example, the techno-utopian narratives of AI-mediated behavioral health "risk detection" at scale to the actual experiences of people described in the subreddit posts of the Reddit dataset in Table 4.

Importantly, included in this dataset aimed at identifying risk levels for suicidal ideation we found a number of individuals who were acutely aware they were depressed and suicidal. They were, in fact, routinely using r/SuicideWatch to express feelings of frustration, despair, anger, and fear, as well as to share their negative experiences with trying to get the support and care they needed from family and medical professionals. People, for example, noted the lack of options for treatment resistant depression, shortcomings of medication regimens, and a lack of financial resources needed to find therapy and alternative forms of treatment, along with widespread lack of trust with clinicians. The posts within this subreddit community as well as others, illustrate common misalignments between the everyday realities of faced by people who have limited access to behavioral healthcare resources and an overly narrow technical vision that motivates research around creating automated "screenings" or predictive risk scores. Furthermore, these experiences surface uncomfortable questions of what kinds of "care" machine learning tools can support when the most pressing patient needs (e.g. affordable access to quality medical care, emotional support, and a safe environment) are tied to social, political, and economic factors typically viewed as outside the scope of AI research. What happens, for instance, after a Reddit user who has struggled for months to find a psychiatrist who will take their insurance, is flagged as high risk for suicide? How does an automated identification of behavioral health conditions meet the messy realities of local healthcare systems? Such questions are essential to consider *before* machine learning models are widely deployed and shape patient experiences of care during periods of intense vulnerability. They also help CSCW researchers and designers reflect on what types of computational interventions are actually useful to people living with behavioral health conditions, not just as a technical puzzle, but as a personal struggle.

4.2.3 The Twitter Dataset. The third Workshop dataset we analyzed in this paper is a dataset composed of social media posts from Twitter, constructed with the intention of building a benchmark for assessing depression and PTSD from a user's Tweets. Twitter, along with Reddit, is a popular site for collective text corpora used to create behavioral health datasets. The Twitter dataset also represents the active collaboration between academic and industry partners. In addition to the many machine learning and clinical psychology researchers outside of the Workshop and the Collective Challenge who have cited the dataset, the companies and organizations represented amongst the authors' affiliations have (whether in parallel or in partnership) also developed similar initiatives and machine learning models to identify PTSD and depression from Twitter users. In fact, research building upon the dataset has been cited by mainstream news publications, which have highlighted research findings developed by the model itself (unusual in comparison to traditional machine learning benchmarks), showcasing the scope of impact of these models to wade far outside of niche academia [2].

While this type of ML work has gained a lot of positive attention as an example of "AI for good," diving into the actual datasets complicates this idealized view of ML. Social media researchers, for example, have long shown the way such platforms include a multiplicity of diverse online communities, each with different discursive styles, performative aspects of tweeting, and other social norms. For example, scholarly work by Brock (2012) has shown how Twitter's interface and online utility mediates Black users' experience of the app, identifying characteristics that facilitate Twitter as a Black "cultural outlet" [12]. However, ML research rarely addresses the cultural context of the Tweets included within the dataset.

Our close reading of the Twitter dataset reveals the specific practices in which this widely used dataset seemingly collapses socially complex conversations into universal "signals" of specific behavioral health conditions. First, in order to provide additional context to support more robust models, hundreds of Tweets are crawled from each user that self-identifies as either having or being diagnosed with depression or PTSD. These “self-identifying” Tweets are first evaluated by a human annotator to determine if the discussion of a behavioral health condition is genuine (and not a “joke”). Along with supplying these Tweets, creators of these types of datasets often include additional information, such as gender and age-matched controls, due to an acknowledgement that “age and gender play a significant role in many mental health conditions, making certain segments of the population more or less likely to be affected or diagnosed with them” [22]. Given that many Twitter users are themselves anonymous (although many are not), these labels are algorithmically determined by an additional classification tool to avoid bias in labeling throughout the dataset. Table 5 provides an excerpt from the Twitter dataset for one user.

Tweets by anonymized_screen_name **hIKfTszSJk**	**age**	**num_tweets**	**gender**	**condition**
@wvzVZbItu6rVJ61 I was first diagnosed in 1980,after getting shot in 76,so I’ve had a lot of time to learn about PTSD. @Si3NRK4FlB0F @pYgrhOC45i3LuA I’m a PTSD victim with 40 years research behind me claims PTSD can be cured in an afternoon are ludicrous @ftucwDK2Z_pThEs As disabled Vietnam vet I’d say it’s nothing new @hr06NrVNqdy0b4_ @hyRGxbSXo @lXjAeZbj6Vl_6B I’m a vet, we put our lives on the line for the USA in jungles,not 5 * hotels	33.64535853	3000	M	ptsd

Table 5. Excerpt from the Twitter dataset, edited for anonymization. Columns left to right show anonymized tweets, algorithmically estimated age, number of tweets, algorithmically estimated gender, and behavioral health condition (based on information provided by the user in a tweet).

Examining samples of the many thousand of tweets included in the dataset reveal misalignments with the stated purpose of identifying those with PTSD or depression with accuracy and precision and how people actually use Twitter, especially in light of varying norms around self-disclosure and presentation of both users (and bots). This can lead to numerous challenges with respect to the goals for accurate detection. For example, from Table 5, it is apparent that the age classification of a Twitter user in the dataset (e.g. 33-44 years) is most likely inaccurate given that this specific account holder also describes themselves as a US veteran who served in Vietnam (an experience, that if true,would make them significantly older than what had been algorithmically determined). While age and gender-matched controls are included in the benchmark, they are not validated methods on their own and can be highly inaccurate. Therefore, given that the Twitter benchmark’s stated purpose is to enable scientific observations and contributions around age and gender of Twitter users and their mental health, such inaccuracies create additional questions regarding the validity of these correlations and its clinical usefulness.

We also find that this dataset is also biased towards “complete” data from users in order to reduce the “noise” that could prevent models from picking up relevant signals. For example, the Twitter dataset skews towards those who have tweeted much more than 1000 times, for both control and individuals identified as being diagnosed with PTSD or depression. Furthermore, we find little information in the Collective Challenge Twitter dataset documentation that offers explanations of missing or incomplete examples that are included in the initial training sets.

Moreover, the Twitter dataset also surfaces important questions around how potential bot activity is accounted for in the construction of the training and test sets. For example, several users within the Twitter dataset exhibit activity that might be considered automated, such as relatively high frequency of tweeting or retweeting the same handles or usernames over a short period of time. One explanation for this is that the dataset specifically captures Twitter metadata that is representative of an older platform, and that this might appear like bot activity in comparison to how Twitter captures a retweet in its current user interface [50]. Given the prevalence of bot activity on social media and growing movements aimed at spreading online health misinformation, research in the space of AI-mediated behavioral health needs to explore both meaningful and misleading signals of behavioral health, and consider how current dataset construction practices might make the diverse particularities of online social worlds invisible.

These construction misalignments further carry on into the vision of how such a model might be applied to address individual and population health concerns. PTSD and depression were specifically chosen as the subject for the Workshop task because of the “high prevalence” of these conditions on Twitter [22]. This shows how the specificity of social media platform, and the widespread desire to use public corpora for machine learning tasks, can influence which behavioral health concerns are targeted as behavioral health problems. While there have been published work in critical data studies on how social media models can increase surveillance and exacerbate privacy concerns of users, we found little discussion in the Workshop participants and the wider ML research community around the social impacts of identifying or characterizing the representative demographics or communities that might be disproportionately represented by high social media activity for the purposes of behavioral health signal detection. And conversely, on who is left out, given that people with some behavioral health conditions are much more likely to consume than produce content related to their conditions [66]. The Twitter dataset, with its datacentric approach to behavioral health detection through access to public data also fails to address the multitude of private ways that people might discuss their lives or show support for one another on social media around sensitive health conditions [54].

4.3 Situating Dataset Protections and Harms

While the social impacts of AI technologies are not always addressed with nuance and care in the broader ML community, it is noteworthy that the Workshop’s organizers have created a number of protections to address possible harms from sharing data and using ML techniques on behavioral health information. The documentation from all three datasets, for instance, addresses user privacy and safety concerns as the primary ethical considerations in dataset construction. For both social media datasets discussed previously, the creators tried to anonymize all metadata that could potentially identify the user, including usernames and URLs. As well, creators of the School Essay dataset undertook anonymization efforts to protect the identities of the individuals populating the datasets. In order for our research team to access a copy of each dataset for our study, we were required to submit proof of IRB approval from our university, along with a statement of intent and a signed letter indicating agreement with the dataset’s terms of use. The Reddit dataset creators also took additional steps to protect privacy by making the dataset available to the American Association of Suicidology (AAS) in order to provide domain expert oversight and accountability.

Privacy concerns also dominated the discourse around the ethical deployment of algorithms that are trained using these datasets. We did not find this discussion within the academic ML research papers that used these datasets; instead these issues were discussed in associated dataset documentation files, as well as on various Collective Challenge websites made by Workshop organizers. These documents provide insight into how computational linguists participating in AI-mediated health understand the ethical landscape of this research space. For example, researchers in this field, such as Coppersmith et al. [23] reflect on the ethical dilemma of deploying algorithmic interventions based on mental health signals derived from social media data. "The crux," they write, is the “trade-off between a person’s right to privacy and the widely agreed-upon moral imperative to act on information that saves lives.” Opt-in principles are described as being an optimal solution to both deploying such algorithms on those who do not wish to be surveilled and those who wish to benefit from “state-of-the-art” screening technologies [23].

We found, however, that such personal design choices around privacy often run counter to the imagined futures of machine learning algorithms in the behavioral health space. Collective Challenge documentation, for instance, makes note of the value of obtaining social media data, as this data represented periods of time when people were in-between doctor visits (what some prominent researchers in this field have termed the “clinical whitespace.”) Training examples generated by individuals who were selected based on self-reporting their behavioral health diagnosis were viewed as limited in their utility, but social media data offered the promise of analyzing data generated by someone who may not be aware that they are, in fact, being monitored for mental health signals. Coppersmith et al. [23] supports this line of reasoning further by stating that “patients cannot always be relied upon to disclose suicidal thoughts in the clinical setting,” emphasizing the need for identifying more implicit linguistic signals that could be used to notify clinicians of a patient’s risk score without necessarily also alerting the user themselves. Following from this perceived need for implicit signal detection, another ethical concern was the impact that false positives might have on the healthcare system as a whole. For example, the notes from one Collective Challenge describe a potential challenge of surveilling the clinical whitespace as the possibility of false positives signals (recommending or intervening in cases that might not be someone truly suffering from suicidal ideation or other mental health issues), “or even true positives” may be detected at a high rate and could ultimately lead to an “overwhelming number of new cases requiring intervention” [88].

These types of ethical framings highlight that one of the primary incentives around dataset construction in ML research communities like the Workshop is the fast-tracking of the usual scientific processes undertaken for clinical algorithms and medical devices through the use of machine learning. Dataset authors argue that the best way to understand the trade-offs and consequences of integrating ML into behavioral health is to build, predict, and involve clinicians along the way. The desire to use such datasets to speed up the discovery process is further evidenced by researchers in the field who maintain that the Healthcare Insurance Portability and Accountability Act (HIPAA), enacted to protect patient privacy of medical data, has left the clinical NLP research far behind the state of the art AI work being done in other domains.

We also note the active industry involvement within the Workshop and the domain of AI-mediated behavioral health, but specifically with respect to the construction of the datasets we discussed in this paper. Companies like Microsoft, Amazon, Facebook and IBM are active in this research field, both sponsoring events, providing research infrastructure, and supporting employee participation. In light of how machine learning science might be used to fast-track behavioral health algorithms not only towards clinical deployment but also across online platforms, we see how corporate incentives might sit uneasily with goals of protecting the privacy of people living with behavioral health conditions.

In summary, the popularity of data science events like the Collective Challenge and the long-term impact of training datasets through benchmarking help define the standards of scientific validity in an emerging field, as well as create public legitimacy for the AI/ML project of detecting and predicting mental health conditions from people's personal data. While these data practices are still very much exploratory research in the AI/ML field–presented as an interesting technical problem for research colleagues to puzzle over—the techniques they develop are actively being used to determine socially complex behavioral health conditions across formal and informal health settings, from military health records in the United State's Veteran's Association (VA) healthcare system to suicide prevention call centers to global social media platforms. Furthermore, as our findings demonstrate, such narrow technical approaches are often misaligned with the lived experience of behavioral health in problematic ways. Next, we turn to discussing the significance of our study findings for critical CSCW health systems research.

5 DISORDERING DATASETS: CRITICAL REFLECTIONS FOR CSCW AI-HEALTH RESEARCH

CSCW/HCI and related fields like Science and Technology Studies (STS) have long explored the challenges around how computational systems meet the complexity of our social worlds [5, 79]. There is a rich literature, for instance, documenting the unintended consequences of technology adoption across healthcare settings [27], as well as the social and emotional impacts of health technologies on people marginalized by traditional design processes [52]. Recently, there is a growing interest in CSCW in understanding not only in how we can better design for AI-Human interaction in healthcare settings, but also in how we as a research community can investigate and address the possible harms AI technologies have for different patient communities [17].

Adding to this nascent critical discourse, in this section we draw together technofeminist and critical writings on AI/ML systems in order to articulate the broader scope of social harms connected to the project of AI-mediated behavioral health. As well, we seek to reflect upon our own position as CSCW researchers and designers by addressing (what is for us) at times a problematic and even distressing technology. As a means of productively troubling the dominant logics of AI/ML systems in behavioral health, we propose the sensitizing concept of *disordering datasets* as a research and design strategy for deliberately turning the analytic lens away from the detection/prediction of disease in patients and back upon the practices of data science itself. Looking at the multiple ways AI/ML datasets can distort, abstract, and problematize the lived experience of behavioral health can help the CSCW community identify places within the AI system development pipeline in need of critical scholarly attention. Furthermore, as a reflexive practice, intentionally situating oneself as part of the social dimensions of AI-mediated health helps make visible the researcher's own methodological blind spots and design commitments [32].

The types of common ML dataset practices exemplified by the Workshop highlight the harms that can come through "pure" technical research work that is isolated from the social messiness of everyday life. In this paper, we identified several sociotechnical misalignments in AI-mediated behavioral health. Key findings from our study include: (1) the misalignments between the historical dimensions of health information and an often context-free approach to ML data collection; (2) the misalignments between people's localized experiences with broken healthcare systems and the idealized visions of AI-mediated behavioral health as a design solution to systemic health inequities; and (3) the misalignments between the sociocultural situatedness of social media communications, and the project of creating clean datasets that represnt a "clinical whitespace.".

The School Essay dataset, for example, illustrated the potential harms that come with ignoring the historical and temporal dimensions of datasets. When data are divorced from their historical context and the lived experience of stigma and oppression in behavioral healthcare is sidelined, there is a real risk that AI systems will perpetuate health injustices and problematically categorize minoritized people as “disordered”or "other." The current AI/ML data practices also assume that a person’s experience of behavioral health follows a fixed linear path through their life, and that it makes sense to use data from decades ago to understand that trajectory. In the Reddit dataset, we drew attention to harms that could occur when data are represented in ways which do not honor people’s situated expertise (see: [77]) of navigating health and healthcare systems. The Reddit users were well-aware of their behavioral health issues–many had unsuccessfully (or with limited success) sought help and had been failed by healthcare institutions. Rather than having a system prioritize them based on calculated “severity,” they would have benefited from solutions that would not legitimize and perpetuate broken healthcare infrastructures and inadequate health policies. Lastly, we saw that the construction of the Twitter dataset made assumptions about users’ ages and genders, as well as assumptions that situated the user outside of any particular online or offline communities. Similar to the School Essay dataset, the application of the Twitter dataset does not take into account the impacts of quickly changing norms, memes, and humor that categorizes much of the communication on social media platforms, or the ways changes in platform design (or current events) might influence how people talk about behavioral health online. Next, we turn to feminist and critical data scholarship to reflect on ways of re-framing AI-mediated behavioral health in ways that center these important historical, infrastructural and socio-cultural dimensions of people’s lived health experiences.

5.1 "Deleting the Social" in AI-Behavioral Health Datasets

In her classic ethnographic studies of early AI systems in medical informatics, Diana Forsythe writes of how AI "deletes the social," a theoretical framing she articulated to highlight the social worlds and situated work of AI researchers and scientists [40]. While ML techniques represents new developments in the state of the art, following Forsythe we too have found AI dataset practices delete the social with respect to reducing complex illnesses like depression, schizophrenia, and bipolar disorder into machine readable signals that aim to provide generalizable insights into behavioral health. Specifically, our analysis calls attention to how such datasets fail to account for the cultural particularities of health information about specific populations, including the wide-ranging social norms that impact what and how people communicate with regards to their illness experience or life events, be it in local school settings or online forums. As well, we found those AI researchers who used these datasets did not account for the wider healthcare environment in which these technologies would (theoretically) one day be deployed. Importantly, as Forsythe argued, such deletions are never value free, but reflect the worldview of those who build and design these systems. The specific deletions that happen in the creation of datasets, for instance, allows for the "detection" and "prediction" of behavioral health conditions to remain a useful and desirable goal, despite most people living in a world in which they have limited ability to address their most immediate and pressing healthcare needs, let alone be in the privileged position to act on information (hopefully?) useful at some future date. A primary contribution of this paper, then, has been to read the social back into these datasets by critically examining the logics, motivations, and politics of the social worlds they are formed within.

In doing so, we propose that ML datasets do not just delete, but also actively disorder the social experience of illness by privileging a biomedical view of disease and operationalizing a shallow and (potentially) harmful algorithmic form of care connected to an engineering ethos of system building and data optimization. Our findings suggest that the Workshop participants, and indeed the wider ML community, frames behavioral health in terms of signal/noise, but equally important to their conceptualization is the Diagnostic and Statistical Manual of Mental Disorders (or DSM-5). The DSM-5 is viewed by clinicians, public health workers, and policy makers as a definitive source for the classification of different mental health conditions, and has been taken up uncritically in machine learning research. The concept of "disorder," however, has been long criticized by patient groups, disability activists, and social scientists as a pejorative term, historically used to pathologize underrepresented and marginalized groups [35]. Along with the social words of particular technical communities, as CSCW researchers and designers we need to actively consider how such medical legacies (along with their associated values) are taken up within the current AI-behavioral health research and in other illness contexts where AI is being applied.

A primary focus in the social worlds developing ML techniques for behavioral health, for instance, is to identify and predict types of behavioral health conditions by detecting patterns in people's personal data. The goal of analyzing personal data to get an ever more precise diagnosis for individual patients reflects a particular understanding of behavioral health based on mathematical theories of information signals and statistical probabilities. Importantly, it also reflects a biomedical view of behavioral health that is often far removed from lived experiences of illness that require people to navigate complex cultural issues such as stigma and identity, as well as systemic barriers like access to a local psychiatric specialist, or the myriad clinical challenges in finding a medication regimen that works.

In the next section, we propose a useful way forward for concerned CSCW researchers and designers who are struggling with how to address issues like the social harms of AI in their work. Turning a critical lens back upon those communities and fields creating AI systems helps make visible the wider context of AI-mediated health as a sociotechnical design space, but also challenges CSCW researchers and designers to actively question the underlying concepts and narratives which can reinforce the technological solutionism found in many AI/ML health applications.

5.2 Disordering Datasets

We propose the analytic framing of 'disordering datasets' as both a critical provocation and research/design strategy for unpacking the wider context of AI-mediated behavioral health (both the social worlds of people living with behavioral health conditions, and the social worlds of those who create AI systems).

The concept of "disorder" as noted in the previous section has historically been used to describe mental illness, and importantly, carries with it negative connotations that reflect the ongoing stigma experienced by those living with behavioral health conditions. Colloquially, the Oxford English dictionary defines *disorder* similarly as "a disturbance of the bodily (or mental) functions; an ailment, disease," but importantly, it also notes that as a verb "disorder" means: "to put out of order; to destroy the regular arrangement of; to throw into disorder or confusion; to disarrange, derange, upset." While machine learning for mental health detection is often held up as an example of "AI for good;" as others have argued, looked at from a different perspective (and social world), these same techniques can also be tools for encoding stigma and a form of digital surveillance [36]. From a patient's point of view, then, AI/ML techniques might be seen as profoundly upsetting in deleting the social side of illness, and throwing into confusion our expectations for care.

We see value as others have done, especially disability advocates, in actively disrupting and questioning who has the power to act in disordered ways. In her book, *Bipolar Expeditions*, anthropologist Emily Martin chronicles stories of people "living under the description of bipolar disorder" who push back on those in power who define rational/irrational behavior [58]. Martin (2007) writes, "Although the form of DSM-IV categories...would seem to speak for their unambiguousness and clarity, in practice they are anything but. Nor do psychiatrists who have the authority to apply these terms to other people always find the process straightforward. What the terms mean, how they should be applied, and even whether the doctor or patient will get to apply them are all matters of contention" [58]. So too, AI/ML approaches to defining behavioral health through data are far from straightforward.

In reclaiming the power of who can disorder, we aim to actively examine the technical work of creating a machine learning training dataset as a social practice. In doing so, we have called attention in this paper to the patterns of practice, design logics, and social realities that don't fit neatly together-e.g., the "mythology and mess" of computing–in the context of AI-mediated health [33]. The analytic lens of *disordering datasets* can also be seen as a technofeminist intervention that reorients our view of AI-mediated behavioral health from a neutral technical process to an assemblage of human-nonhuman actors, including social media platforms, national funding bodies, academic workshops, hack-a-thons, children's old school essays, and subdeddits, etc. This type of analytic upset helps disrupt the seamless view of technological solutionism that dominates how we popularly envision AI/ML applications in the health domain. Importantly, it is also an analytic lens that makes space for diverse (and conflicting) values and lived experiences, including both of people living with behavioral health conditions and the expertise of researchers, engineers, clinicians, and designers working in CSCW, data science or medicine. We turn now to exploring what this can look like in practice.

5.3 Reflexivity as Feminist Praxis in Interdisciplinary Data Science

> *A critical technical practice will, at least for the foreseeable future, require a split identity – one foot planted in the craft work of design and the other foot planted in the reflexive work of critique. Successfully spanning these borderlands, bridging the disparate sites of practice that computer work brings uncomfortably together, will require a historical understanding of the institutions and methods of the field, and it will draw on this understanding as a resource in choosing problems, evaluating solutions, diagnosing difficulties, and motivating alternative proposals.* – Agre [7]

As an interdisciplinary team with a wide range of personal and professional experiences both with AI and behavioral health, we each individually approached this study with varied sensibilities and sensitivities. Engaging in the practice of reflexivity within our situational analysis of AI-mediated behavioral health helped us make sense of how our perspectives fit together (and sometimes remained in tension with one another). In the following section, we explain how reflexivity facilitates the valuable reflective work necessary for understanding complex and sensitive health sociotechical contexts, like ML behavioral health. We also share parts of our own reflexive research process to offer the CSCW community an example of critical and interdisciplinary AI-health research. Given that our design imaginations are often constrained by disciplinary training, interdisciplinary reflexivity can help us to work towards re(making) care as it *should be*—and not continue to design within the world's flawed systems re-enforcing health inequities *as it is*. Here then, is one example of doing AI-health research otherwise.

Ever since the "reflexive turn" in the social sciences, the importance of reflexivity has been widely acknowledged in qualitative research paradigms; however, it is not a method typically used in other disciplinary traditions (especially AI), in part because researchers in those traditions may dismiss reflexivity as "unscientific" [38]. However, STS scholar Donna Haraway argues that recognizing the situatedness of knowledges and taking multiple partial perspectives into account leads to a feminist scientific praxis—in other words, our research is more rigorous when we acknowledge that knowledge is always shaped by our own experiences, communities, and identities [47]. Reflexivity can help CSCW researchers and designers to reflect on the "gap between the possible worlds and the realized world" and develop more imaginative and radical research [76].

Reflexivity, in our project became a necessary component for successful interdisciplinary AI work because of its powerful translational function. It helps to prevent "unidirectional focusing of the discussion or of suggested options for action" into a disciplinary perspective [69]. Venkatasubramanian et al. argue that "By the time individuals [particularly in computer science] have completed disciplinary training, they may already be trapped in structures that do not enable outreach or may even have narrowed perspectives"—in other words, an individual's disciplinary training may prevent them from successfully conducting interdisciplinary research on AI's impact on society [85].

Reflexivity helps us to move beyond computational ways of thinking, as it brings the social front and center instead of "deleting" it. We must create a practice of "design (of technologies, of software, of code) that proceeds from an acknowledgment of our messy entanglements with matter and with each other" [61]—just as we cannot "delete the social," we must recognize the importance of the materiality of systems and their technical implementations. To do so the CSCW/HCI communities will require interdisciplinary research groups that can get "closer to the metal" [13] by analyzing socio-technical implementations and their intertwined values and imaginaries [55]. In other words, McPherson argues that rather than adding the social onto technical research, to do feminist research is to take an interdisciplinary approach from the beginning and carry our commitments throughout our research. As articulated in Drouhard's work [34], collaborative structured reflections, such as the ones we describe in the following section, can help researchers to identify how their research practices relate to these commitments and to navigate the tensions that can arise from bringing together different personal and disciplinary perspectives.

5.3.1 Reflections from our own reflexive research process. Reflexivity can be challenging. It can often seem at first like a vague personal declaration that starts and ends with listing one's identities in a positionality statement. As Mauthner and Doucet note, *being* reflexive is not enough, and *doing* reflexivity throughout data analysis is not as well-understood as a method [60]. They urge researchers to analyze and be transparent about their choices of epistemological and ontological positions, and research practices (e.g., which literatures they cite). Furthermore, reflexivity can be challenging because it is a vulnerable practice–and one that can be a lot to ask of collaborators, particularly when research is related to sensitive topics. This may lead researchers feeling forced to navigate whether to "out" themselves, place their identity in a fixed category, or tell stories they are not ready to tell [6].

We found that in moving beyond positionality, and placing ourselves within "the situation" of AI-mediated behavioral health, we often generated more questions than answers. These uncertainties, however, enabled us to explore different dimensions of ML datasets, and to question our initial close-readings of them. Rose [70] argues for going beyond "transparent reflexivity" that assumes that agency and power are knowable and can be understood through transparency about the researcher's identities. Instead, she argues that researchers should focus on making visible the uncertainties and gaps in their knowledge. Rose quotes Haraway when advocating for the value of "interpretation, translation, stuttering, and the partly understood" [48]. Rose also argues that we need a complex view of power that acknowledges that power is not just constituted between researcher and researched–the reader also has interpretive power. When engaging in reflexivity, we position ourselves as certain kinds of people [25], and thus we, as researchers, cannot escape engaging in the performativity of "being a researcher" because we are co-constitutive with our research—we are shaping discourse and being shaped by it [15, 44].

One goal of this paper, then, is to begin redrawing the discursive and disciplinary boundaries of AI-mediated health through sharing how our backgrounds, lived experiences, ethical commitments, ontological frameworks, and epistemologies became entangled in our study of ML behavioral health datasets [8]. As a team of researchers both within and outside of computer science and data science, over the course of the project we analyzed our relationships to computing discourses around health, noting frustrations with popular language focused on"fixes" and technical "solutions" to complex health issues. In documenting our reflexive process, we ground our previous dataset analysis within our own situated understandings of behavioral health and the larger narratives around care that we saw emerge from detailing the construction practices of ML datasets.

5.3.2 An interdisciplinary reflexivity statement. This study was a part of the Algorithmic Care Project and undertaken while we worked at AI Now Institute, a research group dedicated to investigating the social impacts of AI. Our team came from different disciplinary backgrounds (together the three of us draw together training from across computer engineering, English literature and religious studies, medical sociology, information science, human-computer interaction and design). We all have different forms of privilege and power associated with those research backgrounds and some of our academic credentials (like technical knowledge) may be more privileged than social science expertise in some settings and not others.

We also carry with us our vulnerabilities–on our team, we are all “junior” researchers in some fashion (i.e., postdoc/non-tenured researcher, PhD student, and non-academic). Our relatively precarious positions within our research communities make it challenging to write about our concerns of lucrative fields like data science, and technologies that are often developed by more senior academics in our fields. But we also feel a collective obligation to do what de Castro Leal, Strohmayer, and Krüger talk about and act as good “critical friends” in our communities [28].

We each also come to our AI research work from different intersectional positions with respect to our race, gender, and class; some are shared experiences, some not, but all impact our relationship to health research, care work, and technology. Our lived experiences with race, gender, and class also have shaped our relationships to the way power manifests in the technology space, and we have struggled with what that means for us and our work.

We have each had different experiences with behavioral health, both personally, among our families, friends, and loved ones, and as parts of broader healthcare systems and biomedical infrastructures. These lived experiences shape our research questions, practices, and sensibilities. For example, we found ourselves wondering if we (or our loved ones) would want such AI-mediated behavioral healthcare technologies; how we or they would feel about being (knowingly or unknowingly) included in such datasets; and whether the narratives that accompany the datasets and technologies we studied matched our everyday experiences living with/caring about/researching behavioral health.

We have all experienced different forms of health inequities and systemic injustices; and those experiences have left a mark on our bodies, families, communities, research and career trajectories. In our lives, we have also experienced generative and creative forms of care collectives and hope for the possibilities of health technology design. But we sit uneasily with the tensions of “designing the world as it is vs. world as it should be.”

These questions are difficult, and we engage the feelings of discomfort they bring up, because we do not pretend to have all the answers. Some of our research questions are difficult to answer or have made us feel like we were having an “identity crisis.” We are tangled up in this space (of health and AI ethics), and while we benefit from it in different ways (e.g., professional visibility/a paycheck for our labor), we are also wary of being extractive. Experience with behavioral health in our own lives or those we care about, makes us slow down and try to be careful and intentional in our analysis and research outputs. In analysis sessions, we struggled with the framing of “disordering” datasets and, regardless of our intent, we worried it could be interpreted by others as legitimizing a problematic concept. We, therefore, send our analysis out gently into our research communities, aiming to offer one way of seeing AI-mediated health that we have found useful, but acknowledging many other sensitizing concepts coming from different voices, such as patient communities, are still needed.

Finally, we see ourselves in the datasets, our communities in the uncounted, and our concerns echoed by engineers building the systems we are studying. We wonder about the value of seeing things differently from those participating in the Collective Challenge. And also what it means to find out we might share similar goals? Like those who participate in the Collective Challenge, we too are are fueled by good intentions and desire to make healthcare better. We too see value in "doing the work" of research and of learning how to see from another framing, discipline, or lens. The data scientists and clinicians who we studied have thought deeply about what keeps them up at night and what motivates their work–but what "keeps us all up at night" as CSCW researchers? What is our "intervention" in this space, and what does it mean to push back against technosolutionst and positivist views of (for some) timely and much-needed health research?

We, as a CSCW community, are tasked with the ongoing project of continually and reflexively grappling with these questions. There are no easy answers, and there is no single way forward for CSCW/HCI and adjacent fields that do system building and design work in this domain. But in order to avoid harmful sociotechnical misalignments in AI-Health systems, future work must be attuned to making visible the situations and needs of the people and communities most impacted. We need research that recognizes the many ways the social experience of illness can be deleted or "disordered" by overly narrow and purely technical approaches to dataset construction and conceptualizing AI-health systems.

6 CONCLUSION

Can we identify the (potentially harmful) ways current data practices and the underlying cultural logics of AI-mediated behavioral health will play out in patient communities? Our study points to the limits of a single disciplinary perspective in answering this question, but also the hopeful possibilities in new forms of collaborative work. Epistemologies bound with computation need to be put in balance with social perspectives, and this interdisciplinary work (of which CSCW has a long research tradition) will require reflexivity, vulnerability, and a willingness to learn from one other.

To that end, in this paper, we conducted an interdisciplinary "close reading" of three training datasets used for behavioral health AI systems in order to better understand what technical limitations and social concerns are (in)visible in these datasets. We identified a lack of sociohistorical context for both the data and the classifications used to structure the data; a lack of acknowledgment of biases against minoritized people; gaps in understanding how user data reflects the limitations of many healthcare systems; and missing cultural context of how a platform is used by different groups and used over different versions of the platform with changing norms and affordances.

We found that ethical considerations sometimes became sanitized in the discourse around these datasets or were couched as purely concerns about privacy and surveillance even though much more is at stake. When people's lived experiences do not match how an AI system frames their social worlds and when its paradigms force users to be a part of processes that do not fit with their experiences or needs, there will likely be harmful repercussions, including being denied access to healthcare resources. Challenging aspects of behavioral health (e.g. inadequate insurance coverage, people's inability to pay for medications, and a lack of qualified local behavioral healthcare providers), are too often smoothed over as "merely" social concerns that can be sorted out "later" after the foundational exploratory science is complete and AI systems are rolled out into communities in beta, only to be refined incrementally with live data.

Some CSCW/HCI work situates these problems in such a way that we may be able to build "enlightened" versions of AI-health applications if we just involve the right stakeholders, but our work also suggests that we may not be able to design our way out of every "misalignment" discussed in this paper with traditional HCI methods like participatory design. In fact, in some cases where AI systems are harming patient communities, active resistance and not designing or undesigning may be the only way forward [9, 24]. The reality of computer science is that problems have to be made computationally tractable [10], but the cost may be too high when how humans view problems and how software frames problems are fundamentally different. We must also ask ourselves as researchers (and as a community) what design spaces we should pull out of, what imperfect solutions are acceptable to us, and where we can make long-term changes instead of perpetuating broken systems.

We, as a CSCW research community, are tasked with the ongoing project of continually and reflexively grappling with such grand design challenges [81]. There are no easy answers, and there is not just a single way forward for CSCW and adjacent fields to "fix" ML dataset construction or make every AI-healthcare system equitable. But we can actively seek to better understand the unique sets of sociotechnical misalignments for different health contexts, learn how to "read the social" back into technical processes, and make visible the wider social impacts of ML datasets on diverse illness communities. As Neff et al. [63] argues, data can be an “opportunity to make transparent the assumptions and deliberations that go into choices and to ask more questions, get more input, and build even richer “context”” between people from different disciplinary perspectives and from multiple communities–and it is our hope that our work can help prompt those conversations.

ACKNOWLEDGMENTS

We thank the AI Now Institute for their generous support of the Algorithmic Care Project, and are particularly grateful to Casey Gollan, Alejandro Calcaño Bertorelli, Mariah Peebles, Meredith Whittaker, and Julia Thibaud for their many efforts to support this work. Research reported in this publication was also supported by the Robert Wood Johnson Foundation under award number 77070. The content is solely the responsibility of the authors and does not necessarily represent the official views of the RWJF, AI Now Institute, or other funders.

REFERENCES

[1] [n.d.]. Google Health. https://health.google/for-everyone/self-assessments/

[2] 2014. Twitter opens a window on depression and PTSD - The Boston Globe. https://www.bostonglobe.com/opinion/editorials/2014/08/19/twitter-opens-window-depression-and-ptsd/0elz33EV1dXVwXYVpQ1eTL/story.html

[3] 2021. *Testing times.* Technical Report. National Child Development Study. https://ncds.info/wp-content/uploads/2021/03/NCDS-Booklet-2021-web.pdf

[4] J Khadijah Abdurahman. 2021. Calculating the Souls of Black Folk: Predictive Analytics in the New York City Administration for Children's Services. *Columbia Journal of Race and Law* 11, 4 (2021), 75–110.

[5] Mark S. Ackerman. 2000. The Intellectual Challenge of CSCW: The Gap Between Social Requirements and Technical Feasibility. *Human–Computer Interaction* 15, 2-3 (Sept. 2000), 179–203. https://doi.org/10.1207/S15327051HCI1523_5 Publisher: Taylor & Francis _eprint: https://doi.org/10.1207/S15327051HCI1523_5.

[6] Tony E. Adams and Stacy Holman Jones. 2011. Telling Stories: Reflexivity, Queer Theory, and Autoethnography. *Cultural Studies <-> Critical Methodologies* 11, 2 (April 2011), 108–116. https://doi.org/10.1177/1532708611401329 Publisher: SAGE Publications.

[7] Philip E. Agre. 1997. Toward a Critical Technical Practice: Lessons learned in trying to reform AI. *Social science, technical systems, and cooperative work: Beyond the Great Divide* 131 (1997).

[8] Karen Barad. 2007. *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning.* Duke University Press. Google-Books-ID: H41WUfTU2CMC.

[9] Solon Barocas, Asia J. Biega, Benjamin Fish, Jędrzej Niklas, and Luke Stark. 2020. When not to design, build, or deploy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20).* Association for Computing Machinery, New York, NY, USA, 695. https://doi.org/10.1145/3351095.3375691

[10] Eric P.S. Baumer and M. Six Silberman. 2011. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY,

USA, 2271–2274. https://doi.org/10.1145/1978942.1979275

[11] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan. 2017. Signal Processing and Machine Learning for Mental Health Research and Clinical Applications [Perspectives]. *IEEE Signal Processing Magazine* 34, 5 (Sept. 2017), 196–195. https://doi.org/10.1109/MSP.2017.2718581 Conference Name: IEEE Signal Processing Magazine.

[12] André Brock. 2012. From the blackhand side: Twitter as a cultural conversation. *Journal of Broadcasting & Electronic Media* 56, 4 (2012), 529–549.

[13] Finn Brunton and Gabriella Coleman. 2014. Closer to the Metal. *Media Technologies: Essays on Communication, Materiality, and Society* (2014). https://nyuscholars.nyu.edu/en/publications/closer-to-the-metal Publisher: MIT Press.

[14] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html ISSN: 2640-3498.

[15] Judith Butler. 1999. *Gender trouble: feminism and the subversion of identity.* Routledge, New York.

[16] Scott Allen Cambo and Darren Gergle. 2022. Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science. In *CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3491102.3501998

[17] 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.

[18] Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio, and Munmun De Choudhury. 2019. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 79–88. https://doi.org/10.1145/3287560.3287587

[19] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1201–1213. https://doi.org/10.1145/2818048.2819963

[20] Chelsea Chandler, Peter W Foltz, and Brita Elvevåg. 2020. Using Machine Learning in Psychiatry: The Need to Establish a Framework That Nurtures Trustworthiness. *Schizophrenia Bulletin* 46, 1 (Jan. 2020), 11–14. https://doi.org/10.1093/schbul/sbz105

[21] Adele E. Clarke, Carrie Friese, and Rachel S. Washburn. 2017. *Situational Analysis: Grounded Theory After the Interpretive Turn.* SAGE Publications.

[22] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 Shared Task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Denver, Colorado, 31–39. https://doi.org/10.3115/v1/W15-1204

[23] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights* 10 (Aug. 2018), 1178222618792860. https://doi.org/10.1177/1178222618792860

[24] Sasha Costanza-Chock. 2020. *Design Justice : Community-Led Practices to Build the Worlds We Need.* The MIT Press. https://library.oapen.org/handle/20.500.12657/43542 Accepted: 2020-12-15T13:38:22Z Journal Abbreviation: Community-Led Practices to Build the Worlds We Need.

[25] Bronwyn Davies, Jenny Browne, Susanne Gannon, Eileen Honan, Cath Laws, Babette Mueller-Rockstroh, and Eva Bendix Petersen. 2004. The Ambivalent Practices of Reflexivity. *Qualitative Inquiry* 10, 3 (June 2004), 360–389. https://doi.org/10.1177/1077800403257638 Publisher: SAGE Publications Inc.

[26] Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. 2018. Ethics and artificial intelligence: suicide prevention on Facebook. *Philosophy & Technology* 31, 4 (2018), 669–684. https://doi.org/10.1007/s13347-018-0336-0 Publisher: Springer.

[27] Bas de Boer and Olya Kudina. 2022. What is morally at stake when using algorithms to make medical diagnoses? Expanding the discussion beyond risks and harms. *Theoretical Medicine and Bioethics* (2022), 1–22.

[28] Debora de Castro Leal, Angelika Strohmayer, and Max Krüger. 2021. On Activism and Academia: Reflecting Together and Sharing Experiences Among Critical Friends. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Number 303. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445263

[29] Munmun De Choudhury. 2014. Can social media help us reason about mental health?. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. Association for Computing Machinery, New York, NY, USA, 1243–1244. https://doi.org/10.1145/2567948.2580064

[30] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 3267–3276. https://doi.org/10.1145/2470654.2466447

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

[32] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets. *arXiv:2007.07399 [cs]* (July 2020). http://arxiv.org/abs/2007.07399 arXiv: 2007.07399.

[33] Paul Dourish and Genevieve Bell. 2011. *Divining a Digital Future: Mess and Mythology in Ubiquitous Computing*. MIT Press, Cambridge, MA, USA.

[34] Margaret Drouhard. 2021. *Community Safety Together: How Reflection and Radical Imagination Can Help Us Build the Worlds We Need.* Ph.D. Dissertation. University of Washington.

[35] Steven Epstein. 1996. *Impure science: AIDS, activism, and the politics of knowledge.* Univ of California Press.

[36] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Publishing Group.

[37] Jessica L. Feuston and Anne Marie Piper. 2018. Beyond the coded gaze: Analyzing expression of mental health and illness on instagram. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 51. https://doi.org/10.1145/3274320 Publisher: Association for Computing Machinery (ACM).

[38] Linda Finlay. 2002. "Outing" the Researcher: The Provenance, Process, and Practice of Reflexivity. *Qualitative Health Research* 12, 4 (April 2002), 531–545. https://doi.org/10.1177/104973202129120052 Publisher: SAGE Publications Inc.

[39] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017). https://doi.org/10.2196/mental.7785

[40] Diana E. Forsythe. 2001. *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence.* Stanford University Press, Stanford.

[41] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[42] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *arXiv:1803.09010 [cs]* (March 2020). http://arxiv.org/abs/1803.09010 arXiv: 1803.09010.

[43] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 325–336. https://doi.org/10.1145/3351095.3372862

[44] J. K. Gibson-Graham. 1994. 'Stuffed if I know!': Reflections on post-modern feminist social research. *Gender, Place & Culture* 1, 2 (Sept. 1994), 205–224. https://doi.org/10.1080/09663699408721210

[45] Nazli Goharian, Philip Resnik, Andrew Yates, Molly Ireland, Kate Niederhoffer, and Rebecca Resnik (Eds.). 2021. *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access.* Association for Computational Linguistics, Online. https://aclanthology.org/2021.clpsych-1.0

[46] Daniel Greene, Anna Lauren Hoffmann, and Luke Stark. 2019. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. *Hawaii International Conference on System Sciences 2019 (HICSS-52)* (Jan. 2019). https://aisel.aisnet.org/hicss-52/dsm/critical_and_ethical_studies/2

[47] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575–599. https://doi.org/10.2307/3178066 Publisher: Feminist Studies, Inc.

[48] Donna Haraway. 2013. *Simians, Cyborgs, and Women: The Reinvention of Nature.* Routledge.

[49] Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the State of Social Media Data for Mental Health Research. *arXiv:2011.05233 [cs]* (April 2021). http://arxiv.org/abs/2011.05233 arXiv: 2011.05233.

[50] Andrew Hutchinson. 2020. Twitter Officially Launches New 'Quote Tweets' Count on Main Tweet Display. https://www.socialmediatoday.com/news/twitter-officially-launches-new-quote-tweets-count-on-main-tweet-display/584466/

[51] Shaoxiong Ji. 2020. *Suicidal Ideation Detection in Online Social Content.* Ph.D. Dissertation. The University of Queensland. https://doi.org/10.13140/RG.2.2.19846.32328/1

[52] Elizabeth Kaziunas, Mark S Ackerman, Silvia Lindtner, and Joyce M Lee. 2017. Caring through data: Attending to the social and emotional experiences of health datafication. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing.* 2260–2272. http://doi.org/10.1145/2998181.2998303

[53] Elizabeth Kaziunas, Michael S. Klinkman, and Mark S. Ackerman. 2019. Precarious Interventions: Designing for Ecologies of Care. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 113:1–113:27.

https://doi.org/10.1145/3359215

[54] Danielle Lottridge, Nazanin Andalibi, Joy Kim, and Jofish Kaye. 2019. "Giving a little'ayyy, I feel ya'to someone's personal post" Performing Support on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–22. http://doi.org/10.1145/3359179

[55] Caitlin Lustig. 2019. Intersecting Imaginaries: Visions of Decentralized Autonomous Systems. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 210:1–210:27. https://doi.org/10.1145/3359312

[56] Veronica Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H. Andrew Schwartz. 2018. CLPsych 2018 Shared Task: Predicting Current and Future Psychological Health from Childhood Essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Association for Computational Linguistics, New Orleans, LA, 37–46. https://doi.org/10.18653/v1/W18-0604

[57] Simon Makin. 2019. The emerging world of digital therapeutics. *Nature* 573, 7775 (Sept. 2019), 106–109. https://doi.org/10.1038/d41586-019-02873-1

[58] Emily Martin. 2007. I Now Pronounce You Manic Depressive. In *Bipolar Expeditions*. Princeton University Press, 99–133.

[59] Emily Martin. 2009. *Bipolar Expeditions*. Princeton University Press. Publication Title: Bipolar Expeditions.

[60] Natasha S. Mauthner and Andrea Doucet. 2003. Reflexive Accounts and Accounts of Reflexivity in Qualitative Data Analysis. *Sociology* 37, 3 (Aug. 2003), 413–431. https://doi.org/10.1177/00380385030373002 Publisher: SAGE Publications Ltd.

[61] Tara McPherson. 2014. Designing for Difference. *differences* 25, 1 (May 2014), 177–188. https://doi.org/10.1215/10407391-2420039

[62] Dan Muriello, Lizzy Donahue, Danny Ben-David, Umut Ozertem, and Reshef Shilon. 2018. Under the hood: Suicide prevention tools powered by AI. https://engineering.fb.com/2018/02/21/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/

[63] Gina Neff, Anissa Tanweer, Brittany Fiore-Gartland, and Laura Osburn. 2017. Critique and contribute: A practice-based framework for improving critical data studies and data science. *Big data* 5, 2 (2017), 85–97. https://doi.org/10.1089/big.2016.0050

[64] Ravi B. Parikh, Kristin A. Linn, Jiali Yan, Matthew L. Maciejewski, Ann-Marie Rosland, Kevin G. Volpp, Peter W. Groeneveld, and Amol S. Navathe. 2021. A machine learning approach to identify distinct subgroups of veterans at risk for hospitalization or death using administrative and electronic health record data. *PLOS ONE* 16, 2 (Feb. 2021). https://doi.org/10.1371/journal.pone.0247203 Publisher: Public Library of Science.

[65] Sungkyu Park, Inyeop Kim, Sang Won Lee, Jaehyun Yoo, Bumseok Jeong, and Meeyoung Cha. 2015. Manifestation of Depression and Loneliness on Social Networks: A Case Study of Young Adults on Facebook. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 557–570. https://doi.org/10.1145/2675133.2675139

[66] Jessica A Pater, Brooke Farrington, Alycia Brown, Lauren E Reining, Tammy Toscos, and Elizabeth D Mynatt. 2019. Exploring indicators of digital self-harm with eating disorder patients: A case study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26. http://doi.org/10.1145/3359186

[67] Sachin R. Pendse, Daniel Nkemelu, Nicola J. Bidwell, Sushrut Jadhav, Soumitra Pathare, Munmun De Choudhury, and Neha Kumar. 2022. From treatment to healing: Envisioning a decolonial digital mental health. In *CHI Conference on Human Factors in Computing Systems*. 1–23. https://doi.org/10.1145/3491102.3501982

[68] Chris Power and Jane Elliott. 2006. Cohort profile: 1958 British birth cohort (National Child Development Study). *International Journal of Epidemiology* 35, 1 (Feb. 2006), 34–41. https://doi.org/10.1093/ije/dyi183

[69] Norma R. A. Romm. 1998. Interdisciplinary Practice as Reflexivity. *Systemic Practice and Action Research* 11, 1 (Feb. 1998), 63–77. https://doi.org/10.1023/A:1022964905762

[70] Gillian Rose. 1997. Situating knowledges: positionality, reflexivities and other tactics. *Progress in Human Geography* 21, 3 (June 1997), 305–320. https://doi.org/10.1191/030913297673302122 Publisher: SAGE Publications Ltd.

[71] Pedro Sanches, Axel Janson, Pavel Karpashevich, Camille Nadal, Chengcheng Qu, Claudia Daudén Roquet, Muhammad Umair, Charles Windlin, Gavin Doherty, Kristina Höök, and Corina Sas. 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*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3290605.3300475

[72] Ranjan Satapathy. 2020. NLP for Social Good — Part I. https://medium.com/lingvo-masino/nlp-for-social-good-part-i-85b2d757bf15

[73] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 317:1–317:37. https://doi.org/10.1145/3476058

[74] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 058:1–058:35. https://doi.org/10.1145/3392866

[75] Peter Shepherd. 2013. *Bristol Social Adjustment Guides at 7 and 11 Years: 1958 National Child Development Study User Guide.* Technical Report. Centre for Longitudinal Studies. https://cls.ucl.ac.uk/wp-content/uploads/2017/07/NCDS-Bristol-Social-Adjustment-Guides-final.pdf

[76] Dragos Simandan. 2019. Beyond Haraway? Addressing constructive criticisms to the 'four epistemic gaps' interpretation of positionality and situated knowledges. *Dialogues in Human Geography* 9, 2 (July 2019), 166–170. https://doi.org/10.1177/2043820619850272 Publisher: SAGE Publications.

[77] Jaime Snyder, Elizabeth Murnane, Caitie Lustig, and Stephen Voida. 2019. Visually Encoding the Lived Experience of Bipolar Disorder. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300363

[78] Susan Leigh Star. 1999. The Ethnography of Infrastructure. *American Behavioral Scientist* 43, 3 (Nov. 1999), 377–391. https://doi.org/10.1177/00027649921955326 Publisher: SAGE Publications Inc.

[79] Lucy Suchman. 2006. *Human-Machine Reconfigurations: Plans and Situated Actions* (2 ed.). Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511808418

[80] Anissa Tanweer, Emily Kalah Gade, P. M. Krafft, and Sarah K. Dreier. 2021. Why the Data Revolution Needs Qualitative Thinking. *Harvard Data Science Review* 3, 3 (July 2021). https://doi.org/10.1162/99608f92.eee0b0da

[81] Jacob Thebault-Spieker, Stevie Chancellor, Michael Ann DeVito, Niloufar Salehi, Alex Leavitt, David Karger, and Katta Spiel. 2021. Do We Fix it or Burn it Down? Towards Practicable Critique at CSCW. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '21)*. Association for Computing Machinery, New York, NY, USA, 234–237. https://doi.org/10.1145/3462204.3483281

[82] 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* 27, 5 (Oct. 2020), 1–53. https://doi.org/10.1145/3398069

[83] John Thuma. 2019. The Data Science Behind Natural Language Processing. https://www.arcadiadata.com/blog/the-data-science-behind-natural-language-processing/

[84] Andreas K. Triantafyllidis and Athanasios Tsanas. 2019. Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature. *Journal of Medical Internet Research* 21, 4 (April 2019). https://doi.org/10.2196/12286

[85] Suresh Venkatasubramanian, Nadya Bliss, Helen Nissenbaum, and Melanie Moses. 2020. Interdisciplinary Approaches to Understanding Artificial Intelligence's Impact on Society. *arXiv:2012.06057 [cs]* (Dec. 2020). http://arxiv.org/abs/2012.06057 arXiv: 2012.06057.

[86] Meredith Whittaker. 2021. The steep cost of capture. *Interactions* 28, 6 (Nov. 2021), 50–55. https://doi.org/10.1145/3488666

[87] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI Now Report 2018.* AI Now Institute at New York University New York.

[88] Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, Minneapolis, Minnesota, 24–33. https://doi.org/10.18653/v1/W19-3003

Received July 2021; revised November 2021; accepted April 2022