Illuminating the Unseen: A Framework for Designing and Mitigating Context-induced Harms in Behavioral Sensing

HAN ZHANG, University of Washington, USA
VEDANT DAS SWAIN, Northeastern University, USA
LEIJIE WANG, University of Washington, USA
NAN GAO, Tsinghua University, China
YILUN SHENG, University of Washington, USA
XUHAI XU, Massachusetts Institute of Technology, USA
FLORA D. SALIM, University of New South Wales, Austrilia
KOUSTUV SAHA, University of Illinois Urbana-Champaign, USA
ANIND K. DEY and JENNIFER MANKOFF, University of Washington, USA

With the advanced ability to capture longitudinal sensed data and model human behavior, behavioral sensing technologies are progressing toward numerous wellbeing applications. However, the widespread use of top-down design approaches, often based on assumptions made by technology builders about user goals, needs, and preferences, can result in a lack of context sensitivity. Such oversights may lead to technologies that do not fully support the diverse needs of users and may even introduce potential harms. In this paper, we highlight two primary areas of potential harm in behavioral sensing technologies: identity-based and situation-based harms. By adopting a theory-driven approach, we propose a framework for identifying and mitigating these harms. To validate this framework, we applied it to two real-world studies of behavioral sensing as tools for systematic evaluation. Our analysis provides empirical evidence of potential harms and demonstrates the framework's effectiveness in identifying and addressing these issues. The insights derived from our evaluations, coupled with the reflection on the framework, contribute both conceptually and practically to the field. Our goal is to guide technology builders in designing more context-sensitive sensing technologies, thereby supporting responsible decision-making in this rapidly evolving field.

CCS Concepts: • **Human-centered computing → Empirical studies in ubiquitous and mobile computing**.

Additional Key Words and Phrases: sensing technologies, harms, wellbeing, context-sensitivity, framework, responsibility

ACM Reference Format:
Han Zhang, Vedant Das Swain, Leijie Wang, Nan Gao, Yilun Sheng, Xuhai Xu, Flora D. Salim, Koustuv Saha, Anind K. Dey, and Jennifer Mankoff. 2024. Illuminating the Unseen: A Framework for Designing and Mitigating Context-induced Harms in Behavioral Sensing. 1, 1 (April 2024), 33 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' addresses: Han Zhang, micohan@cs.washington.edu, University of Washington, USA; Vedant Das Swain, v.dasswain@northeastern.edu, Northeastern University, USA; Leijie Wang, leijiew@cs.washington.edu, University of Washington, USA; Nan Gao, nangao@tsinghua.edu.cn, Tsinghua University, China; Yilun Sheng, ylsheng@cs.washington.edu, University of Washington, USA; Xuhai Xu, xoxu@mit.edu, Massachusetts Institute of Technology, USA; Flora D. Salim, flora.salim@unsw.edu.au, University of New South Wales, Austrilia; Koustuv Saha, ksaha2@illinois.edu, University of Illinois Urbana-Champaign, USA; Anind K. Dey, anind@uw.edu; Jennifer Mankoff, jmankoff@cs.washington.edu, University of Washington, USA.

Manuscript submitted to ACM

1 INTRODUCTION

The rapid evolution of sensing technology has created new opportunities to track and reason about human activities. Behavioral sensing technology, which involves the use of sensing technology to capture longitudinal sensed data, and model and predict human behavior from that data, offers a broad spectrum of applications. These include, but are not limited to, wellbeing monitoring (*e.g.*, mental health prediction [3, 92]), human activity recognition (*e.g.*, identifying activities like "running", "sitting", and "walking" [71, 95]), and personalized recommendations (*e.g.*, personalized music and taxi charging recommendation systems [75, 115]). This technology, in contrast to the traditional manual approach of using questionnaire-collected data for the same tasks, facilitates continuous, automated, and unobtrusive gathering of *context* [30, 32]. Here, context refers to capturing all information related to the interactions among users, applications, and their environment [36, 37].

However, alongside the potential for context-aware design, concerns about the responsible use of behavioral sensing technology are growing [25]. These worries arise from the prevalent top-down design approach, which typically involves technology builders – a collective of researchers, designers, developers, and engineers – developing tools based on their assumptions of users' goals, needs, or preferences [68, 90]. Consequently, the application of this technology may lead to a lack of sufficient *context sensitivity*[34, 36], resulting from an incomplete understanding or disregard of contextual factors that may not seem directly relevant to its primary purpose from the early phases of design. Such neglect can result in technology that fails to adequately address the diverse and real-world needs of its users and may even introduce potential harms [126].

In this work, we focus on two key areas of potential harms of behavioral sensing technology that are often overlooked due to the lack of context sensitivity. The first area is identity-based harm, which arises from the ignorance of users' diverse backgrounds. A concerning finding is that only 5% of the research published in a leading journal of sensing technology, PACM IMWUT, from 2018 to 2022 addressed identity-based harm to users with sensitive attributes such as gender or race [126]. This oversight is alarming as it can lead to harmful consequences, particularly for marginalized communities who already face societal inequalities [11, 96, 110]. In response, the broader HCI and CSCW communities are increasingly advocating for a *human-centered* design approach [8]. This approach involves deeply engaging stakeholders to thoroughly understand their experiences and concerns about the impacts of machine learning (ML) and artificial intelligence (AI), especially on marginalized populations (*e.g.*, [7, 35, 66, 76]).

The second area is situation-based harm, which arises when sensing technology is applied or deployed in various situations (*e.g.*, technology infrastructures, environmental conditions, and device types). Typically, technology builders often base their data selection on existing literature or their own experiences when selecting data for algorithm training and testing [17, 90]. However, this practice can introduce potential harms to users. For instance, algorithms trained exclusively with data from iOS-based smartphones might exhibit bias against individuals of lower socioeconomic status who predominantly use less expensive Android-based smartphones [63, 93]. In contrast to identity-based harm, which has received more attention in research, situation-based harm is understudied. This relative lack of study makes such harm more subtle and challenging to identify.

While prior research has shed light on identity-based harm in behavioral sensing [126], providing valuable insights into specific user experiences and perceptions, there remains a significant gap in quantitative research that could offer empirical evidence on a larger scale. Furthermore, we recognize an absence of systematic frameworks that offer clear guidance for technology builders. Such frameworks are essential for designing context-sensitive behavioral sensing technologies that effectively identify and mitigate potential identity-based and situation-based harms.

To address these gaps, this study adopts a theory-driven approach, *human-centered* design approach, to develop a framework for designing context-sensitive behavioral sensing technology. To demonstrate the practicality and effectiveness of our framework, we apply the framework as an evaluation tool to two established and contemporary behavioral sensing studies. These studies focus on assessing two real-world behavioral sensing applications in the domain of wellbeing: mental health detection [124], a classification task; and learning engagement prediction [55], a regression task. A key aspect of our evaluation is examining the extent to which previous work in these areas has considered the harms caused by context insensitivity. Additionally, we conduct a quantitative evaluation within each study to present empirical evidence of the potential harms these applications may pose to users. This approach also serves the purpose of assessing the effectiveness of our framework in identifying and mitigating such harms.

Our findings from both evaluations go beyond the qualitative analysis on the context sensitivity gap in current behavioral sensing research practices [126]. Our quantitative evaluations provide concrete evidence of user harm resulting from this oversight. In the first evaluation study, we identified identity-based harm, where algorithms exhibited bias against marginalized groups. The second evaluation study, while not directly uncovering harm to marginalized groups, reveals significant differences in algorithm performance across different situations, which may also harm certain user groups in situations where algorithms underperform. In summary, our contributions are as follows.

- We highlight the significance of integrating *context-sensitive* considerations in the design of behavioral sensing technologies, highlighting two potential harms due to the oversights regarding context-insensitivity in these technologies.
- Adopting a theory-driven approach, we propose a framework for designing context-sensitive behavioral sensing technologies, aiming to identify and mitigate potential harms to users.
- Through applying our developed framework to assess two real-world wellbeing sensing technology studies, we provide quantitative evidence of potential harms these technologies introduced to users. Furthermore, we make our analysis codebase openly accessible for reproducibility[1].

We further offer key insights from our evaluation studies, as well as reflections within and beyond our proposed framework. Our work aims to contribute both conceptually and practically to the field, focusing on the design of more responsible behavioral sensing technologies.

2 BACKGROUND AND RELATED WORK

As behavioral sensing technologies increasingly become a tool for tracking and reasoning about human activities, they present a blend of promising opportunities and potential risks [128]. In this section, we first review behavioral sensing technology, exploring its evolution and current landscape (Section 2.1). We then review a promising application domain of these technologies, specifically in wellbeing prediction (Section 2.2). Following this, we discuss the potential harms associated with these technologies, including both identity-based and situation-based harms, particularly arising from a lack of context-sensitivity (Section 2.3). We conclude this section by reviewing the human-centered design approach, aimed at addressing these harms (Section 2.4). This background forms the basis for proposing our framework.

[1]We will release our codebase at publication.

2.1 Evolution of Behavioral Sensing Technology

Since the late 1990s, researchers in sensing technology have increasingly recognized the importance of enabling computing devices to enhance application performance by incorporating knowledge of the context in which they are used [1, 36, 105]. "Context" in this sense refers to all information related to the interactions among users, applications, and their environment [37]. Alongside this realization, there was growing advocacy for the creation of sensing systems designed to offer information or services that are relevant to the specific tasks of users, a concept known as *context-awareness* [1, 36]. Building upon this foundational concept, research efforts have since concentrated on creating various toolkits and frameworks to capture, infer, and generate context through diverse sensors [37, 46, 78, 104]. These initiatives have evolved to focus on employing these toolkits for passive data collection, aiming to infer human behavior, such as their phone usage, location, sleep, and steps [39]. More recently, the integration of Machine Learning (ML) and Artificial Intelligence (AI) algorithms into these technologies has further transformed this field. Researchers have begun integrating these advanced techniques into sensing technologies, not only for modeling human behavior but also for making predictions [3, 9, 33].

2.2 Wellbeing Predition in Behavioral Sensing

Wellbeing prediction is one of the promising and extensively studied application domains for behavioral sensing technologies. This domain includes various aspects such as predicting mental health [3, 23, 107] and forecasting performance, engagement, as well as productivity [55, 117]. Specifically, in the area of mental health prediction, considerable research has been dedicated to depression prediction. For instance, studies have utilized passively sensed data such as physical activities, phone usage, sleep patterns, and step counts of participants to predict depressive symptoms [23, 118, 121]. Additionally, there have been significant research efforts aimed at understanding mood-related health concerns among students and workers, employing similar types of passively sensed data [74, 87, 92].

In parallel, evaluating performance, engagement, as well as productivity as a facet of mental health-related wellbeing prediction has also attracted considerable attention. Many studies have focused on student populations. For example, Wang *et al.* [117] conducted a study using passively collected data from smartphones and wearables of college students to predict their cumulative GPA. In another work, Ahuja *et al.* [5] developed a classroom sensing system to capture student facial expressions and body gestures from audio and video data. Their approach allowed for the analysis of students' engagement levels based on these sensory inputs. In a different study, Gao *et al.* [53, 55] employed indoor environmental data, such as temperature, humidity, CO2 levels, and sounds, alongside physical activity data, to predict three dimensions of student engagement levels. Beyond the academic setting, behavioral sensing technology is increasingly being used in the corporate sector to assess workplace productivity and employee wellbeing [99]. For example, Mirjafari *et al.* [89] trained machine learning algorithms on sensing data to differentiate performance levels in workplaces, offering insights for workspace optimization and stress management [33].

Status Quo of Behavioral Sensing Framework: While behavioral sensing applications offer significant opportunities to improve and support human wellbeing, there is an underlying framework common to these technologies that could potentially be harmful. Typically, researchers in this field gather data from wearable devices [55, 107, 121], along with smartphone-embedded apps [23, 122] and other sensors [5, 55]. In addition to these data sources, they often use self-reports [92] or manually annotated labels [5] to collect ground truth data. This data collection is followed by the selection of specific datasets for training and testing algorithms, with the final step usually involving the evaluation of algorithm performance against established baselines. This entire process, while methodical, raises concerns about

potential unintended consequences and risks, particularly the potential harms to users due to a lack of thorough consideration of various contexts during the design process, which we will detail in the next subsection.

2.3 Potential Harms in Behavioral Sensing

As reviewed above, prevailing behavioral sensing technologies often employ a top-down design approach, which is predominantly driven by technology builders' assumptions of users' goals, needs, or preferences [17, 90]. However, developing technologies based solely on these assumptions, alongside what is easily possible to sense, without a thorough understanding of the users' diverse backgrounds and the situation in which the technology is used and deployed, can lead these technologies to a lack of *context sensitivity*. In this paper, we expand context sensitivity, traditionally linked to context-awareness [34, 36], and redefine it to highlight the responsible aspects of these technologies regarding diverse user groups and situations.

> *"A technology is **context-sensitive** when it accounts for diverse user backgrounds, needs, and situations of use to provide value to users."*

As behavioral sensing technologies advance toward practical, real-world applications, it becomes increasingly important to ensure that these technologies are context-sensitive. This consideration is crucial to mitigate potential harms to users. Below, we identify two key areas of potential harms that have been largely understudied or overlooked in this field.

2.3.1 Identity-based Harm. As highlighted by Yfantidou *et al.* [126] in their review of sensing technology research from 2018 to 2022, a mere 5% of studies investigated algorithmic harms to users with sensitive attributes. Alarmingly, 90% of these studies limited their focus to only age and race, and primarily relied on accuracy or error metrics for evaluation. This oversight is concerning, particularly for marginalized groups, who face societal inequalities, as extensively documented in psychology and social work research [24, 51, 67, 106]. For example, Hangartner *et al.* [61] found in their study of online recruitment platforms that individuals from immigrant and minority ethnic groups received 4-19% fewer recruiter contacts. Similarly, Blaser *et al.* [15] noted the significant absence of disability reporting in tech companies and their media coverage. Moreover, Erete *et al.* [44] employed autoethnography[43] and testimonial authority [27] to share their experiences as Black women academics during a pandemic disproportionately impacting their communities and in the context of civil unrest due to racial injustice. When behavioral sensing technology is designed without considering the backgrounds of these marginalized groups, it risks exacerbating existing inequalities, particularly in technologies aimed at predicting mental wellbeing.

2.3.2 Situation-based Harm. Another potential harm that could emerge due to a lack of context sensitivity in behavioral sensing technology is what we identify as situation-based harm. This type of harm could occur when sensing technology is implemented in diverse situations or settings. As this aspect of potential harm is relatively under-explored, we provide an example to help readers conceptually understand it. Specifically, if a behavioral sensing algorithm is predominantly based on data from iOS-based smartphones, it may not be effective on Android-based smartphones due to representational bias [88]. This may potentially lead to a disproportionate impact on individuals of lower socioeconomic status or those in developing countries who commonly use more affordable Android devices. Reportedly, iOS-based smartphones tend to be more than twice as expensive as their Android-based counterparts [63, 93].

2.4 Human-Centered Approach

To identify and mitigate potential harms to humans, there is an increasing call within the HCI and CSCW communities for a focus on *human-centered AI* (HCAI) design [2, 8, 28, 73, 111, 116]. While definitions of HCAI vary [21], the central theme revolves around designing AI technologies that are deeply attuned to the needs, values, and agency of human users, partners, and operators [21]. This approach focuses on ensuring that AI systems are technically efficient and align with human-centric values and ethical standards. Research efforts have been made to provide guidelines for designers in creating more ethical AI systems. For example, Amershi *et al.* [8] distilled over 150 AI-related design recommendations into 18 broadly applicable guidelines, including ensuring AI systems are cautious about social biases and enabling users to provide feedback during regular interaction with the AI systems. Adopting this approach, researchers have actively engaged with stakeholders to gain an in-depth understanding of their experiences and perceptions regarding AI in general and sensing technology in particular. This engagement has focused on various aspects, including stakeholders' trust in AI systems [10, 70, 82, 112], their privacy concerns related to the use of sensing technologies [4, 100, 102], and the specific impacts of these technologies in different settings [25, 31, 32, 68].

Value-sensitive design is another widely adopted approach that can address potential harms to humans [48]. This approach is grounded in the principle of integrating human values into the design process thoroughly and systematically. It utilizes a tripartite methodology that is both integrative and iterative, involving conceptual, empirical, and technical investigations [49, 50]. To explicate this approach, Friedman *et al.* [50] presented three case studies in their work. Leveraging this concept, Zhu *et al.* [129] applied value-sensitive design to algorithm development. In their approach, they actively engaged with stakeholders during the early stages of algorithm creation, incorporating their values, knowledge, and insights to ensure that the technology aligns with user needs and ethical standards.

Drawing from the above extensive body of research, our work integrates a *human-centered* approach into the existing framework for designing behavioral sensing technologies. By applying a human-centered lens, we strive to ensure that our designs not only meet technical requirements but also align with the diverse needs and values of users, thereby mitigating potential harms due to the lack of context sensitivity.

3 A FRAMEWORK FOR DESIGNING CONTEXT-SENSITIVE BEHAVIORAL SENSING TECHNOLOGIES

In this section, we share our design perspectives regarding context-sensitive behavioral sensing technologies (Section 3.1) and introduce our proposed framework (Section 3.2).

3.1 Our Design Perspective: Human-Centered and Context-Sensitive AI

Our design approach is grounded in the principles of *human-centered AI* (HCAI) [65, 108] and *value-sensitive* design [50, 129]. This approach advocates for a shift in the current approach of behavioral sensing technology builders, moving away from the conventional practice of relying on predetermined assumptions about user needs and preferences [90]. Rather, it emphasizes a thorough consideration of potential harms to users and actively incorporates their knowledge and perspectives throughout the technology design process. This aims to ensure that the behavior of the developed technologies does not reinforce negative stereotypes or biases against users, in line with the guidelines proposed by Amershi *et al.* [8]. Furthermore, our approach expands upon the existing concept of context sensitivity in behavioral sensing technology. We extend it from mere awareness of the context to sensitivity and adaptation to changes within that context. We place a strong emphasis on mitigating the potential harms to users by recognizing and addressing

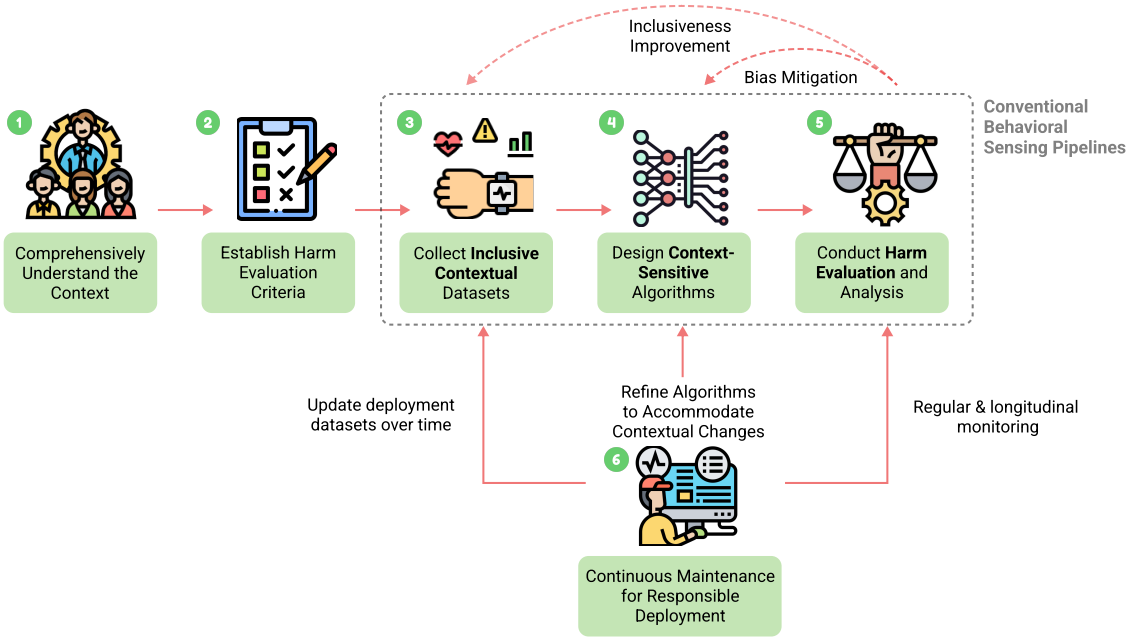


Fig. 1. **Overview of the Framework for Designing Context-Sensitive Behavioral Sensing Technology**. Steps 3, 4, and 5 cover the conventional design flow in the behavioral sensing technology. Note that conventionally, technology builders have often not considered inclusive, context-sensitive, and harm evaluation during data collection, algorithm development, and evaluation (highlighted in **bold** within the dashed box).

both identity-based and situation-based harms that might arise from the technology builders' assumptions regarding the perceived insignificance of certain contexts to the primary objectives (*e.g.*, overall effectiveness) of the technology.

3.2 Overview of the Framework

Inspired by Zhu *et al.* [129]'s work, we first present our developed framework. We then compare our framework with the prevailing design approach in behavioral sensing technology. Our framework, shown in Figure 1, has six steps.

- **Step 1: Comprehensively understand the context.** In this initial phase of developing behavioral sensing technologies, a comprehensive understanding of the *context* is necessary. This involves the awareness of users' diverse backgrounds and engaging with them to understand their specific needs, as well as considering the variety of situated settings (such as technology infrastructure and environmental conditions).
- **Step 2: Establish criteria for evaluating harms, and make sure the bias is not attributed to random choice.** After obtaining a detailed understanding of the context, technology builders should select metrics that can effectively discern algorithmic variances in different contexts. To ensure that these differences are not attributed to random chance, technology builders should employ a rigorous quantitative method, predominantly a statistical analysis [126], for their assessment. The choice of these methods demands careful deliberation to address issues such as Type I errors (false positives) and variations from different groups.

- **Step 3: Collect inclusive datasets including comprehensive contextual information.** This step involves gathering diverse datasets that reflect the comprehensive context understood in Step 1. It is essential to ensure broad representation across a range of demographics and to be acutely aware of situational factors during data collection.
- **Step 4: Develop context-sensitive algorithms.** This step involves engineering algorithms that are sensitive to the potential harms and biases that may emerge from varying contexts. This step requires a careful approach to data selection for training and testing these algorithms. Additionally, technology builders should engage in a continuous cycle of refinement and improvement of these algorithms, particularly when harm or bias is identified.
- **Step 5: Conduct harm evaluation on the behavioral sensing system, combining user feedback and techniques to mitigate harms.** This step involves conducting a comprehensive evaluation of the behavioral sensing technology. The focus here is on identifying and addressing any potential harms or biases that may be present. A key aspect of this evaluation is the integration of user feedback, which provides insights into how the technology performs in real-world scenarios and its impact on different user groups. In instances where biases are observed, technology builders should return to Step 4 to refine the algorithms, combining the insights gained from user feedback.
- **Step 6: Continuous maintenance of data and algorithms for responsible deployment.** This step is critical in the lifecycle of behavioral sensing technology. Once the technology is deployed, technology builders should continuously monitor and update both the data and algorithms and evaluate the model performance to ensure that the technology remains up-to-date and adapts to various contexts.

Comparison with current behavioral sensing technology design approach. In a typical application of behavioral sensing technology for human behavior prediction, the process predominantly revolves around the selection of a specific dataset for model training and validation. This dataset usually consists of input features, derived from sensing data, and a prediction target, which is frequently reliant on self-reported ground truth. The primary objective for the technology builders in this scenario is to develop a model that outperforms the performance of existing models in accuracy or effectiveness (this is also a common theme within the ML community [79]). These steps align only partially with Steps 3, 4, and 5 of our developed framework (as shown in Figure 1).

Our framework, in contrast to this conventional approach, offers a more comprehensive method. Central to our framework is the emphasis on gaining a thorough understanding of the context at the beginning of the technology design process. Beyond recognizing the importance of context, our framework also highlights the need to establish clear evaluation criteria for identifying potential harms. Further, to ensure the validity of findings, we underscore the need for rigorous evaluation to carefully distinguish between genuine biases and anomalies that might arise from random variance. Moreover, our framework emphasizes the ongoing need to maintain and update both data and algorithms to account for the continuous variations in context.

4 EVALUATING EXISTING BEHAVIORAL SENSING TECHNOLOGIES – TWO EVALUATION STUDIES

The objective of this section is two-fold. First, we aim to validate the practicality of our proposed framework. Second, we strive to provide empirical evidence that highlights the potential harms to users that may result from the conventional design approach in behavioral sensing technology.

Table 1. **Overview of Evaluation Criteria and Methods for Each Evaluation Study**. This table enumerates the various elements assessed in each evaluation study and lists the specific methods employed for evaluating the algorithms.

Framework Outline	Evaluation 1: Depression Detection	Evaluation 2: Engagement Prediction
Step 1: Comprehensively understand the context.	**Identity-based harms**: gender, sexual orientation, race, immigration status, first-generation college student status, and disability status. **Situation-based bias**: device types and data collection time. **Engagement with users**: understand their concerns and insights on depression detection sensing technologies.	**Identity-based harms**: gender, disability status, homeless youth, and religious minorities **Situation-based bias**: Temperature condition, location, and class group. **Engagement with users**: understand their concerns and insights on student engagement prediction sensing technologies.
Step 2: Establish criteria for evaluating harms, and make sure the bias is not attributed to random choice.	**Fairness metrics**: disparities in accuracy, false negative rate, and false positive rate. **Significance test**: Mann-Whitney U test with Benjamini-Hochberg correction.	**Fairness metrics**: disparities in mean squared error. **Significance test**: Linear mixed model.
Step 3: Collect inclusive datasets including comprehensive contextual information.	Consider collecting data that could introduce identity-based harms and/or Situation-based bias due to the contextual factors identified in Step 1.	
Step 4: Develop context-sensitive algorithms.	Ensure that the algorithms are aware of harms to users and can adapt to contextual changes.	
Step 5: Evaluate the behavioral sensing, combining user feedback and techniques to mitigate harms.	Assess algorithms for potential harms or biases and verify their mitigation with user feedback.	
Step 6: Continuous maintenance of data and algorithms for responsible deployment.	Implement strategies and actions to regularly update and maintain the data and algorithms.	

To achieve these goals, we apply our framework as a systematic tool to evaluate two real-world behavioral wellbeing sensing technology studies, each within a different domain and involving a distinct ML task. The first study focuses on classifying students with depressive symptoms (Section 4.1). The second study aims to regress students' engagement levels (Section 4.2).

For each evaluation, we start with a background section, delineating the real-world problem, the datasets used, as well as the ML task and algorithms chosen for our evaluation. We then describe our evaluation process (summarized in Table 1). This is followed by the evaluation results. Our evaluations mainly focus on two aspects.

- Evaluate the extent to which the steps proposed in our framework have been considered in previous efforts in the design and implementation of wellbeing sensing technologies.
- Identify the potential harms and biases these technologies might introduce to users, by performing a quantitative evaluation of those algorithms.

We provide detailed descriptions of the evaluation process for steps 1 and 2 within the respective sections of each evaluation study, as these steps are customized to each specific case. Additionally, to gain a deeper understanding of algorithmic harms, we conduct an experiment focusing on bias mitigation in each evaluation study.

4.1 Evaluation Study 1: Depression Detection

Research has been conducted using longitudinal passive sensing data from smartphones and wearable devices to predict and detect depression (*e.g.*, [113, 118, 122]). However, these studies often face challenges related to the limited access to datasets and algorithms, hindering reproducibility and transparency in the field. To address these issues, Xu *et al.* introduced GLOBEM [123], an open-sourced benchmark platform that includes implementations of nine depression detection algorithms and ten domain generalization algorithms. All depression detection algorithms focus on a common binary classification task: distinguishing whether users had at least mild depressive symptoms. They also released a four-year longitudinal passive sensing behavioral dataset from college students [124], aimed to highlight challenges in generalizing and reproducing longitudinal behavior models. In our evaluation study, we examine these depression

Table 2. **Reproduction Results**. Balanced accuracy of the nine depression prediction algorithms on four datasets: DS1 and DS2, which were used in prior work [123], and DS3 and DS4, which are newly reported in this study. The comparison of algorithm performance on DS1 and DS2 ensures the reliability of our fairness evaluation. The Δ column represents the difference between our reproduced results and the previously reported results.

Algorithms	DS1 (2018)			DS2 (2019)			DS3 (2020)	DS4 (2021)
	Prior results	Our results	Diff (Δ)	Prior results	Our results	Diff (Δ)	Our results	Our results
Wahle *et al.* [113]	0.526	0.538	0.012	0.527	0.518	-0.009	0.514	0.514
Saeb *et al.* [103]	0.539	0.539	0.000	0.508	0.513	0.005	0.588	0.500
Farhan *et al.* [45]	0.552	0.552	0.000	0.609	0.609	0.000	0.563	0.609
Canzian *et al.* [20]	0.559	0.538	-0.021	0.516	0.516	0.000	0.541	0.502
Wang *et al.* [118]	0.566	0.565	-0.001	0.500	0.500	0.000	0.577	0.516
Lu *et al.* [80]	0.574	0.574	0.000	0.558	0.558	0.000	0.611	0.553
Xu *et al.*- Interpretable [121]	0.722	0.688	-0.034	0.623	0.667	0.044	0.833	0.733
Xu *et al.*- Personalized [122]	0.723	0.753	0.030	0.699	0.690	-0.009	0.791	0.686
Chikersal *et al.* (removed) [23]	0.728	0.618	-0.110	0.776	0.670	-0.106	0.581	0.641

detection algorithms through a lens focused on potential harms, employing the perspective provided by our proposed framework. Our goal is to assess whether the designs of these algorithms, or their implementations, have considered the steps outlined in our framework. If any of these steps were overlooked, what potential harms can we identify?

Datasets. We chose the four datasets from the GLOBEM study [124] for the evaluation of the depression detection algorithms. To facilitate analysis and comparison, these datasets were labeled chronologically according to the time of their collection (D1 to D4). These four datasets used in our evaluation consist of approximately 700 person-terms of data from around 500 unique participants who were enrolled in the same institution over 10 weeks during the Spring term between 2018 and 2021. These datasets include a wide range of passively sensed behavioral data, including sleep patterns, phone usage statistics, physical activity levels, and phone call records. The datasets also include a wide range of demographics, such as gender, race, first-generation status, immigration status, sexual orientation, and disability status. This data was continuously collected 24 hours per day from smartphones and Fitbits. Additionally, they also include self-reported depression data. We opted to use Beck Depression Inventory-II (BDI-II) scores [12], which were collected once per person at the end of each term in each dataset, as the ground truth.

Depression Detection Algorithms. We chose eight depression detection algorithms implemented by Xu *et al.* [123]. These algorithms consist of a combination of support vector machine [20, 45, 113], logistic regression [103, 118], random forest [113], Adaboost [121], multi-task learning [80], and collaborative-filtering-based model [122]. We excluded one algorithm in the implementation work developed by Chikersal *et al.* [23] from our evaluation study due to a significant disparity between our reproduced results and the reported results in the implementation work [123] (shown in Table 2).

4.1.1 Evaluation Methods and Results. In this subsection, we elaborate on the decision-making processes involved in each step of our framework, followed by presenting our evaluation results.

Step 1: Comprehensively understand the context. In the development of behavioral sensing technologies for depression detection, having a nuanced understanding of user diversity is crucial, as certain sub-populations exhibit higher depression rates. Studies indicate an increased prevalence of depression in specific demographics, such as women [6], first-generation college students [64], immigrants [52], non-heterosexual individuals [130], racial minorities [16], and disability status [127]. These findings underscore the importance for depression detection technology builders to be aware of this context – users' sensitive attributes – to avoid societal biases and ensure equitable outcomes. Additionally,

temporal factors and the type of devices used during data collection are crucial elements to consider. Research has indicated that depressive symptoms can fluctuate based on the time of day (*e.g.*, morning vs. evening) [22, 97] and that user behaviors might be affected by the specific settings of their devices [26]. Recognizing these contexts – timing and device types during data collection – is crucial. Such considerations enable technology builders to accurately model and predict depressive symptoms under different conditions, thereby ensuring the technology's adaptability and fairness in diverse data collection scenarios. Furthermore, engaging users early in the technology development process and incorporating their values and feedback is vital for increasing user acceptance and engagement [50, 129]. This value-sensitive and human-centered approach ensures that the technology is not only technically sound but also resonates with the users' needs and preferences[8, 90].

Building upon the above analysis, our evaluation focuses on assessing whether the designs of these depression detection algorithms take into account three critical aspects: the potential for identity-based harm, situation-based harm, and the extent of technology builders' engagement with users to understand their concerns about mental health sensing technologies (summarized in Table 1).

Step 2: Establish criteria for evaluating harms, and make sure the bias is not attributed to random choice. In this step, we evaluate whether prior work established criteria for evaluating potential harms. To facilitate us to thoroughly examine the potential harms introduced by these depression detection algorithms, we set two key evaluation criteria: classification fairness metrics and thresholds for quantifying differences and biases. In the following subsections, we elaborate on the decision-making process that guided our choices of these criteria. Additionally, we detail the experimental implementation in Appendix A.2, for the sake of transparency and to facilitate reproducibility.

Criterion 1: Classification Fairness Metrics. We used three fairness metrics: disparity in accuracy, disparity in false negative rate, and disparity in false positive rate. These metrics were applied to assess algorithm performance across individuals with sensitive attributes and those without sensitive attributes. We intentionally chose not to adopt commonly used fairness metrics such as demographic parity (*e.g.*, [19]), which aim to ensure equal treatment across different groups. This decision was based on prior research findings indicating that individuals with sensitive attributes are more likely to experience depressive symptoms (*e.g.*, [57, 81, 86]). Using demographic parity, which aims for equal rates of predicted depressive symptoms across groups, could conflict with empirical evidence suggesting inherent disparities in depression prevalence. Our dataset analysis confirmed this, showing notably higher depression levels in certain sensitive groups (first-generation college students, immigrants, and non-male students) from 2018 and 2021[2] (see Figure 4 in Appendix). This highlights the critical need for selecting fairness metrics that reflect real-world disparities.

Criterion 2: Threshold for Quantifying Differences and Biases. We further added a criterion: a threshold quantifying differences in algorithmic performances across various groups. We implemented this to mitigate the impact of random variations. For this purpose, we chose established statistical tests, specifically opting for a non-parametric approach, considering the non-normal distribution of the chosen datasets. We utilized the Mann-Whitney U test, a widely recognized method for comparing means between two independent samples, irrespective of their distribution [83, 120]. We further employed the Benjamini-Hochberg (B-H) correction method to manage the Type I error rate associated with multiple comparisons within the same dataset [13]. We set a stringent False Discovery Rate (FDR) threshold at 0.05 [14, 58], ensuring that the rate of false positives is carefully controlled at 5%.

[2]We performed a Mann-Whitney U test with the Benjamini-Hochberg correction for significance testing (more details are in Section 4.1.1).

Steps 3 to 6: Collect inclusive datasets including comprehensive contextual information; develop context-sensitive algorithms; evaluate the behavioral sensing, combining user feedback and techniques to mitigate harms; and continuous maintenance of data and algorithms for responsible deployment. In evaluating steps 3 to 6, we assessed how the existing depression detection sensing technology builders handled several crucial aspects. First, we looked at whether they took into account the potential for identity-based and situation-based harms during data collection, in line with the contextual factors highlighted in Step 1. Second, we examined the design of the algorithms to determine if they were conscious of potential harms or biases and if they could adapt to contextual changes. Third, we evaluated how these technologies mitigated harms identified in their algorithms, particularly focusing on the use of user feedback. Finally, we assessed whether there were effective strategies and actions in place for the regular updating and maintenance of the data and algorithms.

Below, we present our evaluation findings of current depression detection algorithm designs, focusing on how well they align with our proposed framework. We then present the potential harms identified in these algorithms, based on the context identified in Step 1 and the criteria established in Step 2. This is followed by the experiment of bias mitigation and its results.

Table 3. **Evaluation Results for Two Behavioral Sensing Technology Applications**. This table summarizes the assessment of eight depression detection algorithms (third to tenth rows) based on their original publications, and a re-implementation study (eleventh row). The final line evaluates a student engagement prediction model. Symbols indicate consideration levels: ✓ for full consideration, ⋇ for partial consideration, and ✗ for no consideration.

Algorithm Design / Implementation	Comprehensively Understand Context	Establish Harm Evaluation Criteria	Collect Inclusive Contextual Datasets	Design Context-Sensitive Algorithms	Harm Evaluation and Mitigation	Continuous Maintenance
Evaluation Study 1: Depression Detection						
Wahle *et al.*	⋇	✗	✓	✗	✗	✗
Saeb *et al.*	⋇	✗	✗	✗	✗	✗
Farhan *et al.*	⋇	✗	✓	✗	✗	✗
Canzian *et al.*	⋇	✗	✗	✗	✗	✗
Wang *et al.*	⋇	✗	✓	✗	✗	✗
Lu *et al.*	⋇	✗	✓	✗	✗	✗
Xu *et al.*- Interpretable	⋇	✗	✗	✗	✗	✗
Xu *et al.*- Personalized	⋇	✗	✓	✗	✗	✗
Xu *et al.*- Implementation	⋇	✗	✓	✗	✗	✗
Evaluation Study 2: Engagement Prediction						
Gao *et al.*- En-gage	⋇	✗	✓	✗	✗	✗

Evaluation Results. In our review of nine papers related to the design and implementation of eight depression detection algorithms, we observed that none of the prior work discussed potential harms to users, neither of them engaged with users to better understand their needs. All prior work considered identity-based context, *i.e.*, sensitive attributes. However, consistent with previous sensing technology research, most studies only focused on two sensitive attributes: age [45, 80, 113] and gender [45, 80, 118, 122, 123]. A few also considered race [45, 80, 118, 123], but other sensitive attributes were largely overlooked. In terms of situated aspects, while most studies accounted for data collection time, consideration of device types was less common. Importantly, while these studies reported on this context information, many did not disclose the proportion of data pertaining to each, potentially leading to representative issues. More critically, none of the studies established criteria for evaluating potential harms, nor was there evidence of context-sensitive algorithm design or processes for harm evaluation and mitigation, particularly incorporating user feedback during the whole design process. Furthermore, there was a lack of strategies for the regular maintenance and updating of data and algorithms. These findings are summarized in Table 3, providing an overview of our evaluation results.

To assess the possibility of potential harms arising from a lack of context sensitivity in these depression detection algorithms, we carried out a quantitative analysis. Specifically, we leveraged the evaluation criteria defined in Step 2 (in Section 4.1.1) and evaluated the eight depression detection algorithms on the five sensitive attributes identified in Step 1 (gender, first-generation college student status, immigration status, race, and sexual orientation). Note that, disability status was removed due to the small sample size. The results, detailed in Table 4, revealed several insights.

Firstly, we observed biases in all algorithms towards certain sensitive attributes, *i.e.*, their disparities in accuracy, false negative rates, and false positive rates (highlighted in red in Table 3). Notably, algorithms with higher balanced accuracy [121, 125] tended to show fewer biases across these attributes when evaluated with the three fairness metrics. In particular, the algorithm **Xu_interpretable** [121] did not exhibit bias in terms of accuracy and false positive rate disparities.

Another interesting finding was the reduced bias in all algorithms on DS3, the dataset collected at the start of the COVID-19 outbreak in 2020. This suggests that the significant impact of COVID-19 might have overshadowed other sensitive attributes, leading to this pattern of decreased bias. Additionally, we did not see a consistent pattern indicating which algorithms consistently demonstrated fair performance regarding the sensitive attributes.

Additional Experiment on Bias Mitigation. Recognizing the presence of biases in the eight algorithms, we took steps to mitigate these algorithmic biases. Our approach involved an in-processing technique, where sensitive attributes were incorporated into both the training and testing phases [114, 126]. This method allows algorithms to understand and learn from the relationships between sensitive attributes and the target variable (*i.e.*, BDI-II scores). Previous research indicates that such an approach can help diminish discriminatory patterns present in the data, thereby enhancing the fairness of the models across diverse groups. [114, 126]. Our goal in this experiment was not to develop specific fair algorithms or mitigation techniques but rather to demonstrate a method for reducing bias and obtaining new insights.

As an example, we selected **Xu_interpretable** algorithm [121] due to its relatively high detection performance in depression detection (shown in Table 2) and its relatively low level of disparities across three fairness metrics in the four datasets and five sensitive attributes. We focused on mitigating bias related to first-generation college student status, a sensitive attribute where this algorithm showed bias in terms of disparity of false negative rate. We integrated this attribute into both the training and testing phases of the algorithm and re-evaluated the algorithm's performance.

Our evaluation result, as detailed in Table 5, demonstrates the effectiveness of including the status of being a first-generation college student in the training and testing phases to reduce algorithmic harms. This method led to a fair treatment of this particular sensitive attribute across all datasets, evaluated using three different fairness metrics. However, it is worth noting that while this approach improved fairness for first-generation college student status, it seemed to adversely impact fairness concerning other sensitive attributes such as sexual orientation and gender. A more comprehensive discussion of such trade-offs is described in Section 7.

4.2 Evaluation Study 2: Student Engagement Prediction

4.2.1 Background. In recent years, addressing the growing concerns of poor academic performance and student disinterest has led to a heightened interest in understanding student engagement, emotions, and daily behavior. This shift has coincided with significant advances in sensing technology, paving the way for novel methods to unobtrusively monitor and analyze student behavior and mental well-being in educational settings. A significant milestone in this domain is the introduction of the *En-Gage* dataset by Gao *et al.* [53]. This dataset, available at PhysioNet [3], is

[3]The dataset download link: https://physionet.org/content/in-gauge-and-en-gage/1.0.0/

Table 4. **Algorithmic Harm Evaluation Results**. Results of algorithmic harms through the disparity in accuracy, the disparity in false negative rate, and the disparity in false positive rate (without incorporating demographic data into the training and testing process). The results are adjusted p-values by Benjamini-Hochberg correction after the Mann-Whitney U test. Significance is highlighted in red. Acc, Fnr, and Fpr are the abbreviations of the disparity in accuracy, the disparity in false negative rate, and the disparity in false positive rate.

Algorithms	Sensitive Attributes	DS1 (2018)			DS2 (2019)			DS3 (2020)			DS4 (2021)		
		Acc	Fnr	Fpr	Acc	Fnr	Fpr	Acc	Fnr	Fpr	Acc	Fnr	Fpr
Wahle *et al.* [113]	First-gen College Student	0.020	0.030	0.030	0.010	0.050	0.010	0.020	0.030	0.040	0.010	0.050	0.020
	Gender	0.030	0.030	0.030	0.020	0.030	0.030	0.050	0.010	0.020	0.050	0.030	0.050
	Immigration Status	0.040	0.030	0.030	0.040	0.040	0.050	0.010	0.020	0.040	0.040	0.010	0.030
	Race	0.010	0.030	0.030	0.030	0.010	0.020	0.030	0.050	0.030	0.020	0.020	0.010
	Sexual Orientation	0.050	0.030	0.030	0.050	0.020	0.040	0.040	0.040	0.010	0.030	0.040	0.040
Saeb *et al.* [103]	First-gen College Student	0.010	0.030	0.010	0.020	0.030	0.030	0.010	0.050	0.050	0.020	0.030	0.050
	Gender	0.050	0.040	0.050	0.030	0.030	0.030	0.020	0.030	0.020	0.040	0.040	0.030
	Immigration Status	0.020	0.010	0.040	0.010	0.030	0.030	0.040	0.040	0.030	0.050	0.050	0.020
	Race	0.030	0.050	0.020	0.040	0.030	0.030	0.050	0.010	0.010	0.030	0.020	0.010
	Sexual Orientation	0.040	0.020	0.030	0.050	0.030	0.030	0.030	0.020	0.040	0.010	0.010	0.040
Farhan *et al.* [45]	First-gen College Student	0.030	0.020	0.040	0.020	0.030	0.040	0.030	0.040	0.040	0.030	0.010	0.010
	Gender	0.020	0.030	0.030	0.010	0.020	0.010	0.040	0.030	0.050	0.010	0.040	0.030
	Immigration Status	0.040	0.040	0.020	0.040	0.010	0.030	0.050	0.010	0.030	0.050	0.050	0.050
	Race	0.050	0.050	0.010	0.050	0.050	0.050	0.010	0.020	0.010	0.040	0.020	0.040
	Sexual Orientation	0.010	0.010	0.050	0.030	0.040	0.020	0.020	0.050	0.020	0.020	0.030	0.020
Canzian *et al.* [20]	First-gen College Student	0.020	0.030	0.030	0.020	0.010	0.020	0.020	0.020	0.030	0.010	0.030	0.020
	Gender	0.030	0.030	0.030	0.010	0.020	0.010	0.020	0.030	0.020	0.030	0.020	0.030
	Immigration Status	0.040	0.030	0.030	0.030	0.050	0.050	0.050	0.040	0.040	0.050	0.040	0.050
	Race	0.010	0.010	0.030	0.050	0.030	0.040	0.010	0.020	0.010	0.040	0.030	0.040
	Sexual Orientation	0.050	0.030	0.030	0.040	0.030	0.040	0.040	0.050	0.050	0.020	0.050	0.010
Wang *et al.* [118]	First-gen College Student	0.020	0.050	0.040	0.020	0.030	0.030	0.040	0.030	0.050	0.020	0.040	0.030
	Gender	0.040	0.040	0.030	0.030	0.030	0.030	0.050	0.040	0.020	0.040	0.020	0.040
	Immigration Status	0.010	0.020	0.010	0.010	0.030	0.030	0.010	0.010	0.030	0.050	0.010	0.050
	Race	0.030	0.010	0.020	0.040	0.030	0.030	0.020	0.020	0.040	0.030	0.030	0.020
	Sexual Orientation	0.050	0.030	0.050	0.050	0.030	0.030	0.030	0.050	0.010	0.010	0.050	0.010
Lu *et al.* [80]	First-gen College Student	0.010	0.020	0.030	0.010	0.010	0.050	0.040	0.040	0.040	0.050	0.040	0.030
	Gender	0.050	0.030	0.020	0.040	0.020	0.020	0.030	0.040	0.050	0.020	0.030	0.020
	Immigration Status	0.040	0.050	0.040	0.050	0.030	0.040	0.050	0.010	0.020	0.010	0.050	0.010
	Race	0.020	0.010	0.010	0.030	0.040	0.030	0.010	0.020	0.030	0.040	0.010	0.050
	Sexual Orientation	0.030	0.040	0.030	0.020	0.050	0.050	0.020	0.030	0.050	0.030	0.020	0.020
Xu_interpretable *et al.* [121]	First-gen College Student	0.040	0.020	0.050	0.010	0.030	0.010	0.010	0.010	0.020	0.030	0.010	0.030
	Gender	0.020	0.040	0.040	0.020	0.010	0.020	0.040	0.030	0.040	0.010	0.050	0.010
	Immigration Status	0.030	0.010	0.020	0.050	0.050	0.030	0.030	0.020	0.030	0.050	0.030	0.040
	Race	0.010	0.030	0.010	0.030	0.020	0.040	0.020	0.040	0.010	0.040	0.020	0.050
	Sexual Orientation	0.050	0.050	0.030	0.040	0.040	0.050	0.050	0.050	0.050	0.020	0.040	0.020
Xu_personalized *et al.* [122]	First-gen College Student	0.030	0.040	0.020	0.030	0.030	0.040	0.030	0.040	0.020	0.020	0.020	0.010
	Gender	0.010	0.050	0.040	0.040	0.050	0.030	0.050	0.010	0.010	0.050	0.030	0.020
	Immigration Status	0.050	0.020	0.050	0.020	0.040	0.020	0.010	0.030	0.020	0.040	0.050	0.030
	Race	0.020	0.010	0.030	0.010	0.020	0.010	0.020	0.020	0.050	0.030	0.040	0.040
	Sexual Orientation	0.040	0.030	0.010	0.050	0.010	0.050	0.040	0.050	0.040	0.010	0.010	0.050

distinguished as the largest and most diverse dataset in environmental and affect sensing within the educational field, offering unparalleled insights into student engagement patterns and classroom dynamics through a diverse array of sensing technologies.

Dataset. The En-Gage dataset includes a four-week cross-sectional study involving 23 Year-10 students (15–17 years old, 13 female and 10 male) and 6 teachers (33–62 years old, four female and two male) in a mixed-gender K12 private school. It utilizes wearable sensors to collect physiological data and daily surveys to gather information on the participants' thermal comfort (the comfort level of students regarding the perceived temperature at the time), learning engagement, seating locations, and emotions during school hours. An initial online survey was conducted to obtain

Table 5. **Results After Implementing Bias Mitigation Techniques**. Example of algorithmic fairness evaluation results of Xu_Interpretable [121] through the three fairness metrics (with incorporating first-generation college student status into the training and testing process). The first row in each sub-table showcases the result of our evaluation pertaining to the first-generation college student status, subsequent to the incorporation of this sensitive attribute into both the training and testing phases. In the context of q-values obtained before and after the inclusion of first-generation college student status, ♠ represents the former, while ■ represents the latter. Acc, Fnr, and Fpr are the abbreviations of the disparity in accuracy, the disparity in false negative rate, and the disparity in false positive rate.

Fairness Metric	Sensitive Attribute	DS1 (2018)		DS2 (2019)		DS3 (2020)		DS4 (2021)	
		q value ♠	q value ■	q value ♠	q value ■	q value ♠	q value ■	q value ♠	q value ■
Disparity in Acc	**First-gen College Student**	**0.040**	**0.030**	**0.010**	**0.020**	**0.010**	**0.040**	**0.030**	**0.030**
	Gender	0.020	0.010	0.020	0.010	0.040	0.020	0.010	*0.010
	Immigration Status	0.030	0.050	0.050	0.050	0.030	0.030	0.050	0.040
	Race	0.010	0.040	0.030	0.040	0.020	0.010	0.040	0.020
	Sexual Orientation	0.050	0.020	0.040	0.030	0.050	0.050	0.020	0.050
Disparity in Fnr	**First-gen College Student**	**0.020**	**0.040**	**0.030**	**0.020**	***0.010**	**0.030**	**0.010**	**0.020**
	Gender	0.040	0.010	0.010	0.030	0.030	0.050	0.050	0.050
	Immigration Status	0.010	0.030	0.050	0.050	0.020	0.040	0.030	0.040
	Race	0.030	0.020	0.020	0.040	0.040	0.020	0.020	0.030
	Sexual Orientation	0.050	0.050	0.040	0.010	0.050	0.010	0.040	0.010
Disparity in Fpr	**First-gen College Student**	**0.050**	**0.030**	**0.010**	**0.020**	**0.020**	**0.040**	**0.030**	**0.030**
	Gender	0.040	0.010	0.020	0.010	0.040	0.020	0.010	*0.010
	Immigration Status	0.020	0.020	0.030	0.040	0.030	0.050	0.040	0.050
	Race	0.010	0.050	0.040	0.030	0.010	0.010	0.050	0.020
	Sexual Orientation	0.030	0.035	0.040	0.038	0.050	0.027	0.030	0.040

participants' background information, including age, gender, general thermal comfort, and class groups. The dataset reflects the students' organization into different groups (Form group, Math group, and Language group), aiding in tracking their classroom locations. To clarify, students are typically enrolled in courses based on their form group division, except for math courses which are determined by their math group division, and language courses which are determined by their language group division.

Throughout the study, the participants were asked to wear *Empatica E4* wristbands [85] during school time, which capture 3-axis accelerometer readings, electrodermal activity, photoplethysmography (PPG), and skin temperature. They were also asked to complete online surveys three times a day, posted after certain classes. These surveys capture detailed insights into participants' behavioral, emotional, and cognitive engagement, as well as their emotions, thermal comfort and seating locations [54]. In total, the dataset comprises 291 survey responses and 1415.56 hours of physiological data from all participants.

Engagement Prediction Models. We chose the engagement regression model, LightGBM Regressors [94], developed by Gao *et al.* [55]. The regression model is designed to predict student engagement across three dimensions: *emotional*, *cognitive*, and *behavioral* engagement. Emotional engagement evaluates their feelings of belonging and emotional reaction to the educational environment, cognitive engagement assesses their effort to understand complex ideas and skills, and behavioral engagement looks at students' participation in academic and extracurricular activities. The 1 to 5 Likert scale was used for scoring engagement levels, where 1 represents low and 5 high engagement. To predict these multidimensional scores, a variety of features were extracted, including data from wearable devices and weather stations. It is worth noting that, data such as gender, thermal comfort, and class groups, were not used for the engagement prediction.

4.2.2 Evaluation Methods and Results. In this subsection, following the approach used in Subsection 4.1.1, we first explain the decision-making process for each step of our framework, followed by the results of our evaluation.

Step 1: Comprehensively understand the context. Previous research has highlighted that marginalized groups, including women of color, students with disabilities, homeless youth, and religious minority students, often face feelings of alienation and isolation, which can significantly affect their learning engagement [98, 109]. Studies have also pointed out that variables like the perceived temperature and the timing of data collection can impact student engagement [56, 91]. Moreover, the role of social learning spaces, derived from engaging with student participants, has been recognized as a factor contributing to enhanced engagement [84].

Given these insights, our evaluation of student engagement prediction technology focused on whether its design took these contextual factors into account and aligned with the values and experiences of users. An overview of contextual factors we evaluated can be found in Table 1.

Step 2: Establish criteria for evaluating harms, and make sure the bias is not attributed to random choice. Given the regression-based prediction task, we used the disparity in Mean Squared Error (MSE) as our primary fairness metric to identify biases in model performance. MSE, the average of squared discrepancies between predicted and actual values, is widely recognized for assessing regression model accuracy [119]. Additionally, to discern if biases were systematic or due to random variation, and considering repeated measurements from individuals, we adopted a linear mixed model method [72]. This approach involved calculating residuals (differences between actual values and predictions) across various engagement prediction tasks. Subsequently, we utilized a linear mixed model, executed in Python, to examine whether these residuals significantly varied among different groups (*e.g.*, gender and thermal comfort). This statistical method is beneficial for its ability to account for both within-group and between-group variations in the data, thereby offering a deeper insight into the biases present in model performance.

Steps 3 to 6. Our approach to decision-making and evaluation for Steps 3 to 6 in this evaluation study mirrors the method we employed in the first evaluation study, detailed in Section 4.1.1. A comprehensive overview of the criteria and methods used in these steps can also be found in Table 1.

Evaluation Results. In line with the results from our first evaluation study, our examination of relevant papers in this study [53, 55] indicates that researchers did not engage with the users to understand their needs or considered potential harms to users, and only very limited contextual factors were considered. These factors included gender, thermal comfort at the time of data collection, and information about the courses and classrooms that participants were involved in prior to data collection. Additionally, a key observation is that while this contextual data was considered during the data collection phase, it was not actively incorporated into the training and testing phases of their algorithms. Moreover, the researchers did not address the potential harms of their algorithms. They did not establish criteria for evaluating such harms or implement techniques, including student feedback, to mitigate potential biases. Additionally, there was no evidence of strategies for regular maintenance and updates of the data and algorithms.

We carried out a quantitative analysis to assess the potential negative impacts derived from neglecting certain contextual factors. The findings, detailed in Table 6, indicate that specific situated contexts – such as thermal comfort, the group division (*e.g.*, language and math groups), and the courses students were engaged in prior to data collection – significantly influence the performance of the prediction algorithm. For example, as illustrated in Table 6, the algorithm's ability to accurately assess emotional engagement was statistically different between students who were comfortable with the room temperature and those who were not (feeling either too cold or too warm). To delve deeper into this observation, we analyzed the mean squared error (MSE) of the regression algorithm across different levels of thermal comfort. As reported in Table 7, the algorithm showed a notably lower error rate (MSE = 0.631, p = 0.011) when predicting the emotional engagement of students who were comfortable with the temperature, compared to those who

Table 6. **Results of Linear Mixed Models Analysis**. This table displays the results from linear mixed models, focusing on identifying the significance of differences in regression models across diverse contexts within different engagement prediction tasks. Levels of significance are denoted as follows: * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$. For each contextual factor, one group is designated as the reference (or baseline) category, for example, the Female group in Gender. The "Interpret" represents the average effect for the reference group when all other predictors are held at their reference level (for categorical variables).

Contextual Factors	**Model Variables**	**Emotional Engagement**			**Cognitive Engagement**			**Behavioral Engagement**		
		Coef.	Std. Error	P>|z|	Coef.	Std. Error	P>|z|	Coef.	Std. Error	P>|z|
Gender	Intercept	0.006	0.119	0.962	-0.055	0.121	0.651	-0.033	0.105	0.753
	groups [T.Male]	-0.025	0.179	0.891	0.080	0.182	0.659	-0.038	0.157	0.810
	Group Var	0.121	0.072		0.120	0.070		0.076	0.044	
Thermal Comfort	Intercept	-0.140	0.116	0.225	-0.113	0.118	0.338	-0.136	0.113	0.232
	groups [T.No change]	0.286	0.112	***0.011**	0.181	0.119	0.130	0.215	0.117	0.067
	groups [T.Warmer]	-0.117	0.150	0.436	-0.013	0.160	0.937	-0.190	0.156	0.225
	Group Var	0.112	0.067		0.098	0.054		0.081	0.052	
Language Group	Intercept	0.001	0.094	0.991	0.101	0.114	0.377	0.021	0.102	0.836
	groups [T.Room 41]	0.721	0.244	****0.003**	-0.492	0.299	0.100	0.327	0.264	0.215
	groups [T.Room 43]	-0.094	0.184	0.611	-0.213	0.219	0.331	-0.248	0.198	0.210
	groups [T.Room 68]	-0.351	0.198	0.076	-0.181	0.241	0.452	-0.324	0.214	0.129
	Group Var	0.057	0.045		0.097	0.070		0.068	0.052	
Math Group	Intercept	0.194	0.166	0.242	-0.314	0.155	***0.043**	0.125	0.153	0.414
	groups [T.Room 41]	-0.292	0.215	0.175	0.330	0.201	0.101	-0.335	0.198	0.091
	groups [T.Room 43]	-0.251	0.223	0.261	0.500	0.209	0.017	-0.131	0.205	0.524
	Group Var	0.111	0.070		0.086	0.05		0.082	0.055	
Course	Intercept	-0.285	0.237	0.228	-0.376	0.241	0.119	-0.022	0.237	0.926
	groups [T.English]	0.488	0.241	***0.043**	0.396	0.246	0.107	0.290	0.246	0.239
	groups [T.Health]	0.400	0.353	0.257	0.395	0.360	0.273	-0.086	0.360	0.811
	groups [T.Language]	0.075	0.255	0.769	-0.148	0.260	0.569	-0.400	0.260	0.124
	groups [T.Maths]	0.274	0.240	0.253	0.530	0.245	***0.030**	0.035	0.245	0.885
	groups [T.PE]	0.485	0.356	0.173	0.558	0.363	0.125	0.226	0.363	0.532
	groups [T.Politics]	0.174	0.251	0.489	0.212	0.256	0.407	-0.343	0.257	0.182
	groups [T.Science]	0.240	0.262	0.360	0.634	0.267	***0.018**	-0.085	0.267	0.749
	Group Var	0.128	0.085		0.126	0.073		0.084	0.054	

were not (MSE = 0.822 for students feeling the temperature should be cooler and MSE = 0.742 for students feeling the temperature should be warmer). Similarly, our analysis indicated a significantly higher error rate (MSE = 0.849, p = 0.043) in predicting the cognitive engagement of students in Room 40 for their math class, as opposed to those in other math groups (MSE = 0.715 for Room 41 and MSE = 0.753 for Room 43). Interestingly, our analysis revealed no evidence of algorithmic bias or harm, both with gender and in predicting student behavioral engagement.

Additional Experiment on Bias Mitigation. To determine if the findings from our first evaluation study can be replicated using the same bias mitigation technique – incorporate context data into both the training and testing phase of the algorithm – we conducted an additional experiment in this evaluation study.

As an example, we aimed to mitigate the algorithmic bias caused by the lack of detailed information about students' assignments to different language groups, specifically in the context of predicting student emotional engagement. As indicated in Table 8, incorporating this information into both the training and testing phases of the algorithm proved effective in reducing algorithmic harm. Compared to Table 3, this method resulted in a more equitable prediction performance across students assigned to various language groups. However, it was less effective in addressing biases related to different levels of thermal comfort and the variety of courses students were taking.

Table 7. **Overview of Basic Statistics.** MSE refers to the Mean Squared Error, indicating the average of the squares of the errors. 'Residual' denotes the difference between the ground truth and prediction. MR represents the Mean Residual, which is the average of residuals within each group. "Ind" and "Obs" stand for individuals and observations, respectively.

Context Factors	Groups	Counts (Ind/Obs)	Emotional Engagement		Cognitive Engagement		Behavioral Engagement	
			MSE	MR	MSE	MR	MSE	MR
Gender	Female	13/149	0.708	-0.014	0.711	-0.005	0.800	-0.023
	Male	10/142	0.693	0.033	0.822	-0.004	0.674	0.002
Thermal Comfort	No Change	22/163	0.631	0.158	0.774	0.069	0.687	0.129
	Cooler	20/77	0.822	-0.140	0.755	-0.135	0.730	-0.115
	Warmer	14/51	0.742	-0.242	0.751	-0.045	0.915	-0.300
Language Group	Room 40	13/155	0.618	0.008	0.690	0.156	0.681	0.051
	Room 41	2/53	0.886	0.703	1.019	-0.563	0.799	0.427
	Room 43	5/52	0.526	-0.075	0.779	-0.082	0.592	-0.186
	Room 68	3/53	1.007	-0.312	0.823	-0.072	1.016	-0.276
Math Group	Room 40	7/80	0.671	0.247	0.849	-0.327	0.640	0.178
	Room 41	9/110	0.763	-0.114	0.715	0.044	0.783	-0.185
	Room 43	7/101	0.657	-0.046	0.753	0.197	0.768	0.030
Course	Chapel	11/12	0.938	-0.268	1.100	-0.297	0.693	0.021
	English	18/71	0.599	0.255	0.484	0.057	0.639	0.340
	Health	8/8	0.991	0.155	1.538	0.149	0.776	-0.035
	Language	20/38	0.986	-0.206	0.970	-0.512	0.975	-0.372
	Maths	20/79	0.551	-0.033	0.816	0.150	0.779	0.0177
	PE	8/8	1.044	0.235	1.314	0.103	0.845	0.224
	Politics	19/43	0.754	-0.119	0.784	-0.101	0.643	-0.341
	Science	19/32	0.641	0.006	0.537	0.252	0.687	-0.050

5 DISCUSSION

In this section, we begin by summarizing the key insights we derived from our evaluation studies (Section 5.1). This summary covers the various findings, their implications, and how they contribute to our understanding of designing behavioral sensing technologies. Following this, we delve into a reflection on our framework, examining its strengths, limitations, and considering perspectives that extend beyond its current scope (Section 5.2).

5.1 Key Insights on Evaluation Studies

Our evaluations of two real-world behavioral wellbeing sensing technology studies demonstrated the practicality and effectiveness of our proposed framework. Throughout both evaluation studies, we identified a range of commonalities as well as unique findings, which we detail below.

5.1.1 Potential Harms to Marginalized Groups Due to Context-insensitivity. In both of our evaluation studies, we uncovered a critical and consistent issue with existing behavioral sensing technology designs: a widespread disregard for potential harms to users. Our evaluation, as detailed in Table 3, revealed that none of the designs thoroughly considered steps 2, 4, 5, and 6 proposed in our framework during their design processes. Furthermore, while a few designs did consider the collection of more diverse contextual datasets (*e.g.*, [121, 122]), this type of data was not utilized effectively during the algorithm training and testing phases. Our quantitative analysis of algorithm performance substantiates the concern of potential harms due to this oversight. Both studies identified significant issues, either identity-based harm or situation-based harm. Identity-based harm, which is more straightforward, can directly impact marginalized groups. In contrast, the concept of situation-based harm is more nuanced and can be subtle in its impact on these groups. To illustrate this further, in addition to the example discussed in Section 2.3.2, our second evaluation study

Table 8. **Results After Implementing Bias Mitigation Techniques**. The outcomes following the inclusion of language group assignment data in the training and testing phases of the emotional engagement prediction algorithm.

Contextual Factors	Model Variables	Emotional Engagement		
		Coef.	Std. Error	P>|z|
Gender	Intercept	-0.026	0.096	0.785
	groups [T.Male]	0.030	0.143	0.834
	Group Var	0.060	0.045	
Thermal Comfort	Intercept	-0.154	0.103	0.137
	groups [T.No change]	0.288	0.110	**0.009
	groups [T.Warmer]	-0.094	0.147	0.523
	Group Var	0.055	0.042	
Language Group	Intercept	0.022	0.093	0.811
	groups [T.Room 41]	0.236	0.240	0.325
	groups [T.Room 43]	-0.060	0.181	0.740
	groups [T.Room 68]	-0.282	0.195	0.147
	Group Var	0.053	0.043	
Math Room	Intercept	0.164	0.131	0.210
	groups [T.Room 41]	-0.299	0.169	0.077
	groups [T.Room 43]	-0.172	0.175	0.327
	Group Var	0.050	0.042	
Course	Intercept	-0.345	0.228	0.131
	groups [T.English]	0.579	0.240	*0.016
	groups [T.Health]	0.554	0.351	0.115
	groups [T.Language]	0.111	0.254	0.663
	groups [T.Maths]	0.304	0.239	0.203
	groups [T.PE]	0.490	0.354	0.166
	groups [T.Politics]	0.202	0.251	0.420
	groups [T.Science]	0.320	0.261	0.219
	Group Var	0.051	0.040	

provides another insightful instance. Specifically, we found that the algorithm for predicting emotional engagement was less effective for students who felt uncomfortably cold or warm compared to those who were comfortable with the temperature (as detailed in Tables 6 and 7). This finding may imply a potential indirect harm to individuals of lower socioeconomic status, who may have restricted access to air conditioning and thus are more likely to experience algorithmic harms [60].

5.1.2 A Need for Engage Users Throughout the Design Process. Another key finding from both of our evaluation studies is the complete absence of user involvement throughout the design process of existing behavioral sensing technologies. Given the widespread use of behavioral sensing technologies, particularly in the mental health domain, this is concerning. As argued by Zhu *et al.* [129], engaging with users in the early stage of the design process can ensure that technologies are designed with a deep understanding of users' needs and values, which can significantly enhance user acceptance and satisfaction. Furthermore, the engagement of users extends beyond the initial design phase to include ongoing feedback loops. Regular interactions with users allow for iterative improvements and adjustments based on evolving needs, emerging challenges, and changing social contexts [8]. However, it is important to recognize the balance between involving users to mitigate technology harms and minimizing demands on their time and resources. This is especially necessary for people with different needs [127]. Striking this balance ensures that users' contributions are meaningful and sustainable, and that their valuable input genuinely shapes the direction of the technology while respecting their availability and capacity.

5.1.3 Balance Trade-offs in Achieving Algorithmic Fairness. In our first evaluation study, as detailed in Section 4.1.1, we encountered a trade-off when attempting to mitigate harms. We observed that while using an in-processing mechanism

(*i.e.*, incorporating the sensitive attribute into the training and testing process) helped reduce bias for that particular attribute, it unexpectedly introduced a bias towards other sensitive attributes. This outcome highlights the complex and nuanced nature of mitigating algorithmic harms. It suggests that while certain mitigation strategies might address specific biases, they can also unintentionally create new harms.

Understanding these intricate trade-offs, technology builders should explicitly ask the question when developing context-sensitive algorithms: which group of users should be prioritized and to what extent? Our framework emphasized the importance of comprehensively understanding users' backgrounds and specific needs to answer the first part of this question. Our emphasis on engaging more with users and involving them throughout the design of the behavioral sensing technologies answers the latter part of this question.

The complex field of algorithmic fairness often presents trade-offs not only between different user groups but also among various fairness metrics [62]. Each metric provides a unique lens on bias, focusing on different aspects of equity. However, optimizing for one metric may lead to unintended negative outcomes in another, creating challenging scenarios [29]. For example, in our first evaluation study, when sensitive attributes were incorporated into Saeb *et al.*'s algorithm training and testing, it reduced bias in accuracy disparity for most sensitive attributes, as shown in Table 9. Yet, a detailed examination of other key fairness metrics like disparity in false negative rates and positive rates (Tables 10 and 11) reveals significant variations. This highlights the complex dynamics involved in fairness optimization, where achieving fairness in one dimension might inadvertently lead to imbalances in others, emphasizing the multifaceted nature of achieving algorithmic fairness. This dynamic becomes even more crucial in behavioral sensing technologies, where data collection remains continuous, and system behavior is deeply adaptive to changing contexts. This recognition sets the stage for our subsequent discussion (Section 5.2), delving into the essential requirement for regular and systematic monitoring.

5.2 Towards More Responsible Behavioral Sensing

In this section, we discuss various aspects both within and beyond our current framework. These include the necessity of continuous maintenance for longitudinal behavioral sensing deployment while minimizing human labor, considerations of harms in other components of behavioral sensing technology, as well as the incorporation of other responsible considerations. Our intention is to inspire researchers and designers towards the conception and realization of more responsible behavioral sensing technologies.

5.2.1 Need for Regular Maintenance while Alleviating Excessive Human Labor. During our evaluation studies, we discovered that the reason behind the lack of continuous maintenance for responsible deployment (as outlined in step 6 of our framework) in these behavioral sensing technology designs was that the technologies were not truly deployed in real-world settings. This limitation arises from the nature of the limited datasets and the absence of deployable algorithmic systems [123, 128].

Behavioral sensing technologies in real-world applications operate in dynamic environments and depend on continuously evolving data streams. This dynamic nature increases the risk of situation-based harms, demanding continuous vigilance to guarantee that the system's accuracy and fair alignment persist over time. Regular maintenance is a key step in achieving this goal. By continuously updating deployment datasets, refining algorithms to accommodate temporal dynamics, and regularly monitoring the system's performance, the system can uphold its reliability and fulfill its ethical responsibility towards users and users. However, it is also important to avoid overburdening human resources with excessive maintenance demands. High human labor requirements can lead to operational inefficiencies, increased costs,

and hindered scalability [47]. Striking a balance between rigorous maintenance and an approach that minimizes the burden on human resources is pivotal. Leveraging automation, intelligent monitoring, and adaptive algorithms can potentially alleviate this issue.

5.2.2 Other Aspects of Harms in Behavioral Sensing Technology. In this work, our emphasis centers on addressing harms within the algorithmic aspects of behavioral sensing technology. Nonetheless, it is important to acknowledge that considerations of harms should extend beyond the algorithm design process and include other critical components of the system. One dimension, for example, to focus on is the user interface and user interaction.

A user interface that is designed without incorporating the consideration of harms to users can inadvertently influence users to make certain choices or take specific actions. When these nudges disproportionately benefit particular groups, it can result in disparate outcomes that perpetuate inequality. Additionally, if user interfaces are not designed with accessibility in mind, individuals with disabilities might face barriers in accessing the interacting with the system. Future endeavors should take this aspect into account when designing their user interface. Finally, approaches to transparently informing users about potential fairness concerns, similar to transparent information about accuracy concerns, should be incorporated into a deployed fair behavioral sensing technology.

5.2.3 Expanding Responsible Considerations to Address Additional Needs. While our study primarily concentrates on algorithmic harms in the context of behavioral sensing technology, it is essential to recognize that responsible considerations encompass a broader spectrum of dimensions, such as transparency, privacy, and accountability (*e.g.*, [42, 77]). As behavioral sensing technology becomes more widely used, ensuring transparency becomes a pressing concern. A lack of transparency can lead to opacity and a lack of user trust [10, 18]. Interpretability and explanation techniques are crucial in addressing this issue, allowing users to understand algorithmic decisions and aiding system developers in identifying potential harms [40]. Furthermore, continuous data collection in behavioral sensing raises significant privacy challenges [100]. The risk of unauthorized and unintended data sharing is ever-present. Researchers can develop privacy-preserving algorithms and techniques tailored specifically to behavioral sensing environments and delve into privacy-enhancing technologies, such as secure multi-party computation [59, 69], federated learning [101], and differential privacy [41], and incorporate them into the framework. In addition, accountability is about establishing mechanisms to hold responsible parties accountable for the outcomes of their algorithms and systems [38]. In behavioral sensing, accountability can be challenging due to complex decision-making processes and interactions between various components. Technology builders of these technologies must be held answerable for their impact on users. Expanding our proposed framework to include all the above-discussed aspects can create a more comprehensive foundation for responsible behavioral sensing technology design and deployment.

6 LIMITATIONS AND FUTURE WORK

While our research included two comprehensive evaluations of real-world behavioral sensing technologies across various domains and machine ML tasks, aiming to derive broader conclusions, we recognize that both evaluation studies are situated within the overarching theme of wellbeing prediction. This specific focus may limit the generalizability of our findings to other applications outside of wellbeing prediction. Furthermore, in both of our evaluation studies, we identified various instances of identity-based and situation-based harms. However, we note that some aspects might still be overlooked. Moreover, the datasets used in our evaluations presented their own set of constraints. For instance, specific instances of harms, such as disability status, were either unrepresented or underrepresented due to limited data collection or small sample sizes. This data limitation restricts our ability to make conclusive statements about

these groups. Future work should continue to explore the potential harms in a broader array of behavioral sensing technology applications, identify additional instances both within the two potential harms we discussed and beyond, as well as collect more inclusive datasets for a more comprehensive analysis.

7 CONCLUSION

In conclusion, this paper introduces a framework developed to assist technology builders in designing context-sensitive behavioral sensing technologies. Our framework offers a structured approach to address considerations of potential harms due to a lack of context sensitivity. Through our two evaluation studies, we showcase the practical applicability of our proposed framework. By conducting quantitative analyses, we not only uncover empirical evidence of potential harms in existing behavioral sensing technologies but also validate the framework's capability in identifying and mitigating these harms. We discuss the insights learned from the evaluation studies, as well as other aspects within and beyond the scope of our proposed framework. We hope our work inspires technology builders in our field to amplify their attention to the significance of incorporating harm considerations and other responsible considerations in behavioral sensing technologies.

REFERENCES

[1] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles. Towards a better understanding of context and context-awareness. In *Handheld and Ubiquitous Computing: First International Symposium, HUC'99 Karlsruhe, Germany, September 27–29, 1999 Proceedings 1*, pages 304–307. Springer, 1999.

[2] D. Acemoglu. Harms of ai. Technical report, National Bureau of Economic Research, 2021.

[3] D. A. Adler, D. Ben-Zeev, V. W. Tseng, J. M. Kane, R. Brian, A. T. Campbell, M. Hauser, E. A. Scherer, and T. Choudhury. Predicting early warning signs of psychotic relapse from passive sensing data: an approach using encoder-decoder neural networks. *JMIR mHealth and uHealth*, 8(8):e19962, 2020.

[4] D. A. Adler, E. Tseng, K. C. Moon, J. Q. Young, J. M. Kane, E. Moss, D. C. Mohr, and T. Choudhury. Burnout and the quantified workplace: Tensions around personal sensing interventions for stress in resident physicians. *Proceedings of the ACM on Human-computer Interaction*, 6(CSCW2):1–48, 2022.

[5] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and Y. Agarwal. Edusense: Practical classroom sensing at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–26, 2019.

[6] P. R. Albert. Why is depression more prevalent in women? *Journal of psychiatry & neuroscience: JPN*, 40(4):219, 2015.

[7] A. Alkhatib. To live in their utopia: Why algorithmic systems create absurd outcomes. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–9, 2021.

[8] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[9] N. Banovic, T. Buzali, F. Chevalier, J. Mankoff, and A. K. Dey. Modeling and understanding human routine behavior. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 248–260, 2016.

[10] N. Banovic, Z. Yang, A. Ramesh, and A. Liu. Being trustworthy is not enough: How untrustworthy artificial intelligence (ai) can deceive the end-users and gain their trust. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–17, 2023.

[11] S. Barocas, M. Hardt, and A. Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017.

[12] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri. Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of personality assessment*, 67(3):588–597, 1996.

[13] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[14] Y. Benjamini and D. Yekutieli. Quantitative trait loci analysis using the false discovery rate. *Genetics*, 171(2):783–790, 2005.

[15] B. Blaser, C. Bennett, R. E. Ladner, S. E. Burgstahler, J. Mankoff, C. Frieze, and J. Quesenberry. *Perspectives of women with disabilities in computing*. Cambridge, UK: Cambridge Univ. Press, 2019.

[16] J. Breslau, S. Aguilar-Gaxiola, K. S. Kendler, M. Su, D. Williams, and R. C. Kessler. Specifying race-ethnic differences in risk for psychiatric disorder in a usa national sample. *Psychological medicine*, 36(1):57–68, 2006.

[17] M. Brodie, E. Pliner, A. Ho, K. Li, Z. Chen, S. Gandevia, and S. Lord. Big data vs accurate data in health research: large-scale physical activity monitoring, smartphones, wearable devices and risk of unconscious bias. *Medical hypotheses*, 119:32–36, 2018.

[18] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[19] F. Buet-Golfouse and I. Utyagulov. Towards fair unsupervised learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1399–1409, 2022.

[20] L. Canzian and M. Musolesi. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 1293–1304, 2015.

[21] T. Capel and M. Brereton. What is human-centered about human-centered ai? a map of the research landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2023.

[22] I. Chelminski, F. R. Ferraro, T. V. Petros, and J. J. Plaud. An analysis of the "eveningness–morningness" dimension in "depressive" college students. *Journal of affective disorders*, 52(1-3):19–29, 1999.

[23] P. Chikersal, A. Doryab, M. Tumminia, D. K. Villalba, J. M. Dutcher, X. Liu, S. Cohen, K. G. Creswell, J. Mankoff, J. D. Creswell, et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(1):1–41, 2021.

[24] T. Chou, A. Asnaani, and S. G. Hofmann. Perception of racial discrimination and psychopathology across three us ethnic minority groups. *Cultural Diversity and Ethnic Minority Psychology*, 18(1):74, 2012.

[25] S. Chowdhary, A. Kawakami, M. L. Gray, J. Suh, A. Olteanu, and K. Saha. Can workers meaningfully consent to workplace wellbeing technologies? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 569–582, 2023.

[26] K. Church, D. Ferreira, N. Banovic, and K. Lyons. Understanding the challenges of mobile phone usage data. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 504–514, 2015.

[27] P. H. Collins. *Intersectionality as critical social theory*. Duke University Press, 2019.

[28] E. Commission. Ethics guidelines for trustworthy ai. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf, 2019.

[29] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[30] V. P. Cornet and R. J. Holden. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of biomedical informatics*, 77:120–132, 2018.

[31] S. Corvite, K. Roemmich, T. I. Rosenberg, and N. Andalibi. Data subjects' perspectives on emotion artificial intelligence use in the workplace: A relational ethics lens. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.

[32] V. Das Swain, L. Gao, W. A. Wood, S. C. Matli, G. D. Abowd, and M. De Choudhury. Algorithmic power or punishment: Information worker perspectives on passive sensing enabled ai phenotyping of performance and wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.

[33] V. Das Swain, K. Saha, M. D. Reddy, H. Rajvanshy, G. D. Abowd, and M. De Choudhury. Modeling organizational culture with workplace experiences shared on glassdoor. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15, 2020.

[34] N. Davies, K. Mitchell, K. Cheverst, and G. Blair. Developing a context sensitive tourist guide. In *1st workshop on human computer interaction with Mobile devices, GIST technical report G98-1*, volume 1, 1998.

[35] A. DeVos, A. Dhabalia, H. Shen, K. Holstein, and M. Eslami. Toward user-driven algorithm auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[36] A. K. Dey. Understanding and using context. *Personal and ubiquitous computing*, 5:4–7, 2001.

[37] A. K. Dey, G. D. Abowd, and D. Salber. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human–Computer Interaction*, 16(2-4):97–166, 2001.

[38] N. Diakopoulos. Accountability in algorithmic decision making. *Communications of the ACM*, 59(2):56–62, 2016.

[39] A. Doryab, P. Chikarsel, X. Liu, and A. K. Dey. Extraction of behavioral features from smartphone and wearable data. *arXiv preprint arXiv:1812.10394*, 2018.

[40] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[41] C. Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12. Springer, 2006.

[42] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2021.

[43] C. Ellis, T. E. Adams, and A. P. Bochner. Autoethnography: an overview. *Historical social research/Historische sozialforschung*, pages 273–290, 2011.

[44] S. Erete, Y. A. Rankin, and J. O. Thomas. I can't breathe: Reflections from black women in cscw and hci. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–23, 2021.

[45] A. A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. In *2016 IEEE wireless health (WH)*, pages 1–8. IEEE, 2016.

[46] D. Ferreira, V. Kostakos, and A. K. Dey. Aware: mobile context instrumentation framework. *Frontiers in ICT*, 2:6, 2015.

[47] C. B. Frey and M. A. Osborne. The future of employment: How susceptible are jobs to computerisation? *Technological forecasting and social change*, 114:254–280, 2017.

[48] B. Friedman. Value-sensitive design. *interactions*, 3(6):16–23, 1996.

[49] B. Friedman, D. G. Hendry, A. Borning, et al. A survey of value sensitive design methods. *Foundations and Trends® in Human–Computer Interaction*, 11(2):63–125, 2017.

[50] B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren. Value sensitive design and information systems. *Early engagement and new technologies: Opening up the laboratory*, pages 55–95, 2013.

[51] D. M. Frost. Social stigma and its consequences for the socially stigmatized. *Social and Personality Psychology Compass*, 5(11):824–839, 2011.

[52] K. Fung and C.-L. Dennis. Postpartum depression among immigrant women. *Current opinion in Psychiatry*, 23(4):342–348, 2010.

[53] N. Gao, M. Marschall, J. Burry, S. Watkins, and F. D. Salim. Understanding occupants' behaviour, engagement, emotion, and comfort indoors with heterogeneous sensors and wearables. *Scientific Data*, 9(1):261, 2022.

[54] N. Gao, M. S. Rahaman, W. Shao, K. Ji, and F. D. Salim. Individual and group-wise classroom seating experience: Effects on student engagement in different courses. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–23, 2022.

[55] N. Gao, W. Shao, M. S. Rahaman, and F. D. Salim. n-gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–26, 2020.

[56] E. Garrosa, L. M. Blanco-Donoso, I. Carmona-Cobo, and B. Moreno-Jiménez. How do curiosity, meaning in life, and search for meaning predict college students' daily emotional exhaustion and engagement? *Journal of Happiness Studies*, 18:17–40, 2017.

[57] J. L. Givens, T. K. Houston, B. W. Van Voorhees, D. E. Ford, and L. A. Cooper. Ethnicity and preferences for depression treatment. *General hospital psychiatry*, 29(3):182–191, 2007.

[58] M. E. Glickman, S. R. Rao, and M. R. Schultz. False discovery rate control is a recommended alternative to bonferroni-type adjustments in health studies. *Journal of clinical epidemiology*, 67(8):850–857, 2014.

[59] O. Goldreich. Secure multi-party computation. *Manuscript. Preliminary version*, 78(110), 1998.

[60] C. J. Gronlund. Racial and socioeconomic disparities in heat-related health effects and their mechanisms: a review. *Current epidemiology reports*, 1:165–173, 2014.

[61] D. Hangartner, D. Kopp, and M. Siegenthaler. Monitoring hiring discrimination through online recruitment platforms. *Nature*, 589(7843):572–576, 2021.

[62] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[63] M. Jamalova and M. Constantinovits. The comparative study of the relationship between smartphone choice and socio-economic indicators. *International Journal of Marketing Studies*, 11(3):11, 2019.

[64] S. R. Jenkins, A. Belanger, M. L. Connally, A. Boals, and K. M. Durón. First-generation undergraduate students' social support, depression, and life satisfaction. *Journal of College Counseling*, 16(2):129–142, 2013.

[65] M. I. Jordan. Artificial intelligence—the revolution hasn't happened yet. *Harvard Data Science Review*, 1(1):1–9, 2019.

[66] N. Karizat, D. Delmonaco, M. Eslami, and N. Andalibi. Algorithmic folk theories and identity: How tiktok users co-produce knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–44, 2021.

[67] S. Karlsen and J. Y. Nazroo. Relation between racial discrimination, social class, and health among ethnic minority groups. *American journal of public health*, 92(4):624–631, 2002.

[68] A. Kawakami, S. Chowdhary, S. T. Iqbal, Q. V. Liao, A. Olteanu, J. Suh, and K. Saha. Sensing wellbeing in the workplace, why and for whom? envisioning impacts with organizational stakeholders. *arXiv preprint arXiv:2303.06794*, 2023.

[69] M. Keller. Mp-spdz: A versatile framework for multi-party computation. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 1575–1590, 2020.

[70] S. S. Kim, E. A. Watkins, O. Russakovsky, R. Fong, and A. Monroy-Hernández. Humans, ai, and context: Understanding end-users' trust in a real-world computer vision application. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 77–88, 2023.

[71] H. Kwon, C. Tong, H. Haresamudram, Y. Gao, G. D. Abowd, N. D. Lane, and T. Ploetz. Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(3):1–29, 2020.

[72] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[73] M. K. Lee, N. Grgić-Hlača, M. C. Tschantz, R. Binns, A. Weller, M. Carney, and K. Inkpen. Human-centered approaches to fair and responsible ai. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.

[74] B. Li and A. Sano. Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–26, 2020.

[75] J. Li, Z. He, Y. Cui, C. Wang, C. Chen, C. Yu, M. Zhang, Y. Liu, and S. Ma. Towards ubiquitous personalized music recommendation with smart bracelets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–34, 2022.

[76] C. A. Liang, S. A. Munson, and J. A. Kientz. Embracing four tensions in human-computer interaction research with marginalized people. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 28(2):1–47, 2021.

[77] Q. V. Liao and J. W. Vaughan. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941*, 2023.

[78] B. Y. Lim and A. K. Dey. Toolkit to support intelligibility in context-aware applications. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 13–22, 2010.

[79] L. T. Liu, S. Wang, T. Britton, and R. Abebe. Reimagining the machine learning life cycle to improve educational outcomes of students. *Proceedings of the National Academy of Sciences*, 120(9):e2204781120, 2023.

[80] J. Lu, C. Shang, C. Yue, R. Morillo, S. Ware, J. Kamath, A. Bamis, A. Russell, B. Wang, and J. Bi. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–21, 2018.

[81] N. B. Lucero, R. L. Beckstrand, L. C. Callister, and A. C. Sanchez Birkhead. Prevalence of postpartum depression among hispanic immigrant women. *Journal of the American Academy of Nurse Practitioners*, 24(12):726–734, 2012.

[82] S. Ma, Y. Lei, X. Wang, C. Zheng, C. Shi, M. Yin, and X. Ma. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[83] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[84] K. E. Matthews, V. Andrews, and P. Adams. Social learning spaces and student engagement. *Higher Education Research & Development*, 30(2):105–120, 2011.

[85] C. McCarthy, N. Pradhan, C. Redpath, and A. Adler. Validation of the empatica e4 wristband. In *2016 IEEE EMBS international student conference (ISC)*, pages 1–4. IEEE, 2016.

[86] D. L. McFadden. Health and academic success: A look at the challenges of first-generation community college students. *Journal of the American Association of Nurse Practitioners*, 28(4):227–232, 2016.

[87] L. Meegahapola, W. Droz, P. Kun, A. De Götzen, C. Nutakki, S. Diwakar, S. R. Correa, D. Song, H. Xu, M. Bidoglia, et al. Generalization and personalization of mobile sensing-based mood inference models: An analysis of college students in eight countries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–32, 2023.

[88] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[89] S. Mirjafari, K. Masaba, T. Grover, W. Wang, P. Audia, A. T. Campbell, N. V. Chawla, V. D. Swain, M. D. Choudhury, A. K. Dey, et al. Differentiating higher and lower job performers in the workplace using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–24, 2019.

[90] D. C. Mohr, K. R. Weingardt, M. Reddy, and S. M. Schueller. Three problems with current digital mental health research... and three things we can do about them. *Psychiatric services*, 68(5):427–429, 2017.

[91] A. Morrison, S. Rozak, A. Gold, and J. Kay. Quantifying student engagement in learning about climate change using galvanic hand sensors in a controlled educational setting. *Climatic Change*, 159:17–36, 2020.

[92] M. B. Morshed, K. Saha, R. Li, S. K. D'Mello, M. De Choudhury, G. D. Abowd, and T. Plötz. Prediction of mood instability with passive sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–21, 2019.

[93] J. D. Ndibwile, E. T. Luhanga, D. Fall, D. Miyamoto, and Y. Kadobayashi. Smart4gap: Factors that influence smartphone security decisions in developing and developed countries. In *Proceedings of the 2018 10th International Conference on Information Management and Engineering*, pages 5–15, 2018.

[94] M. Nemeth, D. Borkin, and G. Michalconok. The comparison of machine-learning methods xgboost and lightgbm to predict energy development. In *Computational Statistics and Mathematical Modeling Methods in Intelligent Systems: Proceedings of 3rd Computational Methods in Systems and Software 2019, Vol. 2 3*, pages 208–215. Springer, 2019.

[95] L. Peng, L. Chen, Z. Ye, and Y. Zhang. Aroma: A deep multi-task learning based simple and complex human activity recognition method using wearable sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–16, 2018.

[96] D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[97] A. A. Putilov. Associations of depression and seasonality with morning-evening preference: comparison of contributions of its morning and evening components. *Psychiatry research*, 262:609–617, 2018.

[98] S. J. Quaye, S. R. Harper, and S. L. Pendakur. *Student engagement in higher education: Theoretical perspectives and practical approaches for diverse populations*. Routledge, 2019.

[99] M. S. Rahaman, J. Liono, Y. Ren, J. Chan, S. Kudo, T. Rawling, and F. D. Salim. An ambient–physical system to infer concentration in open-plan workplace. *IEEE Internet of Things Journal*, 7(12):11576–11586, 2020.

[100] C. Reynolds and R. Picard. Affective sensors, privacy, and ethical contracts. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1103–1106, 2004.

[101] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

[102] J. Rooksby, A. Morrison, and D. Murray-Rust. Student perspectives on digital phenotyping: The acceptability of using smartphone data to assess mental health. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[103] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, D. C. Mohr, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research*, 17(7):e4273, 2015.

[104] D. Salber, A. K. Dey, and G. D. Abowd. The context toolkit: Aiding the development of context-enabled applications. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 434–441, 1999.

[105] B. Schilit, N. Adams, and R. Want. Context-aware computing applications. In *1994 first workshop on mobile computing systems and applications*, pages 85–90. IEEE, 1994.

[106] M. T. Schmitt and N. R. Branscombe. The meaning and consequences of perceived discrimination in disadvantaged and privileged social groups. *European review of social psychology*, 12(1):167–199, 2002.

[107] Y. S. Sefidgar, W. Seo, K. S. Kuehn, T. Althoff, A. Browning, E. Riskin, P. S. Nurius, A. K. Dey, and J. Mankoff. Passively-sensed behavioral correlates of discrimination events in college students. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–29, 2019.

[108] B. Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[109] A. L. Sjogren and T. N. Melton. The complexities of student engagement for historically marginalized youth in an after-school program. *Journal of Youth Development*, 16(5):105–121, 2021.

[110] H. Suresh and J. V. Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2(8), 2019.

[111] M. Tahaei, M. Constantinides, D. Quercia, S. Kennedy, M. Muller, S. Stumpf, Q. V. Liao, R. Baeza-Yates, L. Aroyo, J. Holbrook, et al. Human-centered responsible artificial intelligence: Current & future trends. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–4, 2023.

[112] K. Vodrahalli, R. Daneshjou, T. Gerstenberg, and J. Zou. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 763–777, 2022.

[113] F. Wahle, T. Kowatsch, E. Fleisch, M. Rufer, S. Weidt, et al. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR mHealth and uHealth*, 4(3):e5960, 2016.

[114] M. Wan, D. Zha, N. Liu, and N. Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):1–27, 2023.

[115] G. Wang, Y. Zhang, Z. Fang, S. Wang, F. Zhang, and D. Zhang. Faircharge: A data-driven fairness-aware charging recommendation system for large-scale electric taxi fleets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–25, 2020.

[116] Q. Wang, M. Madaio, S. Kane, S. Kapania, M. Terry, and L. Wilcox. Designing responsible ai: Adaptations of ux practice to meet responsible ai challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

[117] R. Wang, G. Harari, P. Hao, X. Zhou, and A. T. Campbell. Smartgpa: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 295–306, 2015.

[118] R. Wang, W. Wang, A. DaSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, and A. T. Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–26, 2018.

[119] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.

[120] F. Wilcoxon. *Individual comparisons by ranking methods*. Springer, 1992.

[121] X. Xu, P. Chikersal, A. Doryab, D. K. Villalba, J. M. Dutcher, M. J. Tumminia, T. Althoff, S. Cohen, K. G. Creswell, J. D. Creswell, et al. Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–33, 2019.

[122] X. Xu, P. Chikersal, J. M. Dutcher, Y. S. Sefidgar, W. Seo, M. J. Tumminia, D. K. Villalba, S. Cohen, K. G. Creswell, J. D. Creswell, et al. Leveraging collaborative-filtering for personalized behavior modeling: a case study of depression detection among college students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(1):1–27, 2021.

[123] X. Xu, X. Liu, H. Zhang, W. Wang, S. Nepal, Y. Sefidgar, W. Seo, K. S. Kuehn, J. F. Huckins, M. E. Morris, et al. Globem: Cross-dataset generalization of longitudinal human behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–34, 2023.

[124] X. Xu, H. Zhang, Y. Sefidgar, Y. Ren, X. Liu, W. Seo, J. Brown, K. Kuehn, M. Merrill, P. Nurius, et al. Globem dataset: Multi-year datasets for longitudinal human behavior modeling generalization. *arXiv preprint arXiv:2211.02733*, 2022.

[125] Z. Xu, J. Li, Q. Yao, H. Li, X. Shi, and S. K. Zhou. A survey of fairness in medical image analysis: Concepts, algorithms, evaluations, and challenges. *arXiv preprint arXiv:2209.13177*, 2022.

[126] S. Yfantidou, M. Constantinides, D. Spathis, A. Vakali, D. Quercia, and F. Kawsar. Beyond accuracy: A critical review of fairness in machine learning for mobile and wearable computing. *arXiv preprint arXiv:2303.15585*, 2023.

[127] H. Zhang, M. Morris, P. Nurius, K. Mack, J. Brown, K. Kuehn, Y. Sefidgar, X. Xu, E. Riskin, A. Dey, et al. Impact of online learning in the context of covid-19 on undergraduates with disabilities and mental health concerns. *ACM Transactions on Accessible Computing*, 15(4):1–27, 2022.

[128] H. Zhang, L. Wang, Y. Sheng, X. Xu, J. Mankoff, and A. Dey. A framework for designing ubiquitous computing systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (Adjunct)*, 2023.

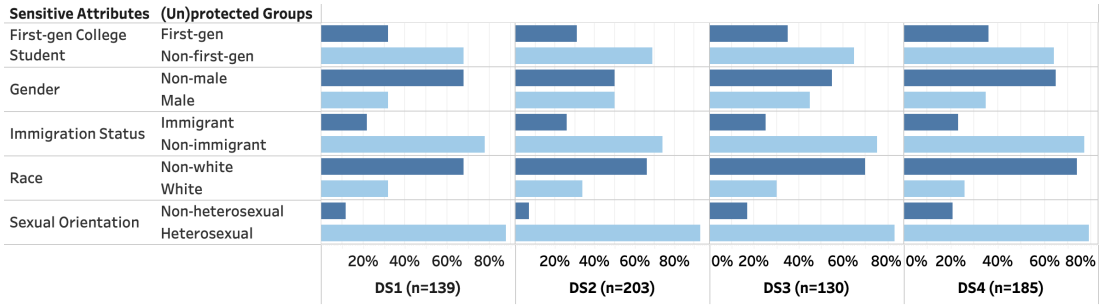
[129] H. Zhu, B. Yu, A. Halfaker, and L. Terveen. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–23, 2018.

[130] B. P. Zietsch, K. J. Verweij, A. C. Heath, P. A. Madden, N. G. Martin, E. C. Nelson, and M. T. Lynskey. Do shared etiological factors contribute to the relationship between sexual orientation and depression? *Psychological Medicine*, 42(3):521–532, 2012.

A APPENDIX

A.1 Percentage of Each Group within Each Sensitive Attribute

Fig. 2. Percentage of each group within each sensitive attribute. The protected group for each sensitive attribute (*e.g.*, first-gen) is shaded in dark colors, while the unprotected group is shaded in light colors (*e.g.*, non-first-gen). Non-male includes women, transgender individuals, and genderqueer individuals, non-heterosexual includes homosexual, bisexual, and asexual individuals, and non-white includes black, asian, latinx, and biracial.

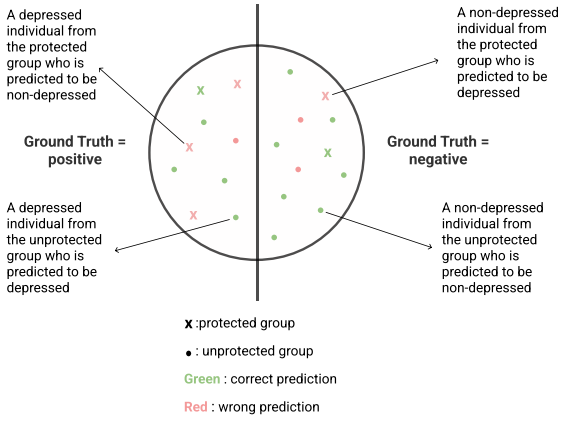


A.2 Case Study 1: Example of the Statistical Evaluation and Experimental Implementation

Fig. 3. Example of fairness evaluation based on the disparities in accuracy, false negative rate, and false positive rate. (a) shows the synthetic data for 20 individuals, with 6 belonging to the protected group (represented by "x" marks) and 14 belonging to the unprotected group (represented by "•" marks). (b) visualizes the distribution and disparities of predictions for both groups, where correction predictions are depicted in green and incorrect predictions in red.

PID	True Label	Prediction	Protected Group	Assigned Value
1	1	0	Yes	0
2	1	1	Yes	1
3	1	1	No	1
4	0	0	No	1
5	0	0	No	1
6	1	1	No	1
7	1	1	No	1
8	0	1	Yes	0
9	0	0	No	1
10	1	0	Yes	0
11	0	1	No	0
12	0	0	Yes	1
13	1	0	Yes	0
14	1	1	No	1
15	0	0	No	1
16	0	0	No	1
17	1	0	No	0
18	0	0	No	1
19	0	1	No	0
20	0	0	No	1

(a) Synthetic data for 20 individuals.



(b) Visualization of prediction distributions and disparities.

A.2.1 Example of the Statistical Evaluation. In Figure 3, we present an illustrative example to demonstrate our approach to fairness evaluation. In this example, we generated synthesized ground-truth labels and predictions from an algorithm for a sample of 20 individuals. Among these individuals, 6 are part of the protected group, while 14 belong to the

unprotected group (as shown in Figure 3a). We assigned a value of "1" for accurate predictions and "0" for inaccurate predictions based on the correctness of the predictions.

Figure 3b visualizes the distribution of predictions for both the protected group (represented by "x") and the unprotected group (represented by "•"). The circle in the figure represents the distribution of predictions, with the left side indicating cases where all ground truth values are positive (representing individuals with depression in our case study), and the right side representing cases where all ground truth values are negative (representing individuals without depression in our case study). The accuracy of the predictions is indicated by the color, with green representing correct predictions by the algorithm and red representing incorrect predictions.

In this example, when considering the disparity in accuracy, the algorithm made incorrect predictions for 4 out of 6 individuals from the protected group (represented by the red "x" marks among all both red and green "x" marks). Conversely, for the unprotected group, the algorithm made incorrect predictions for 3 out of 14 individuals (depicted by the red "•" marks among all red and green "•" marks). To assess the statistical significance of these disparities, we conducted the Mann-Whitney U test in combination with the Benjamini-Hochberg correction. Specifically, we applied this test to the 2 "1" values and 4 "0" values corresponding to the protected group, as well as the 11 "1" values and 3 "0" values corresponding to the unprotected group.

When examining the difference in false negative rates, the relevant information for statistical analysis is contained in the left portion of the circle depicted in Figure 3b. Specifically, we conducted a statistical test on the 1 "1" value and 3 "0" values in the protected group, as well as the 4 "1" values and 1 "0" value in the unprotected group. Similarly, an evaluation of the disparity in false positive rates was conducted on the marks on the right side of the circle in Figure 3b.

A.2.2 Experimental Implementation. We applied the two evaluation criteria as defined in Section 4.1.1 to evaluate the fairness of the eight depression detection algorithms. We provide a detailed explanation of our statistical analyses to capture disparities in accuracy, false negative rate, and false positive rate below (an example of this approach can be found in above). We provide open access to our evaluation codebase to enable reference and reproducibility for future research.

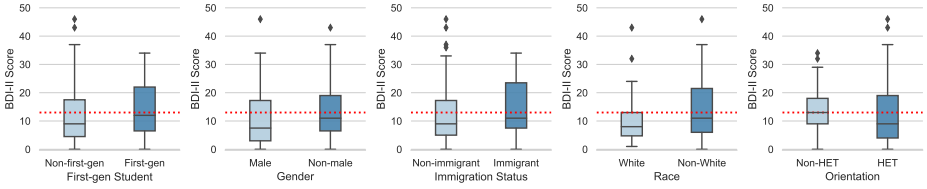
To perform the Benjamini-Hochberg correction, we first calculated the p values for all attributes using the Mann-Whitney U test. Then, we arranged the p values in ascending order and assign ranks to them, with the smallest p value receiving rank 1, the second smallest receiving rank 2, and so on. Next, we calculated the adjusted q values for each individual p value using the formula: $(i/m) \times Q$, where i is the rank of the individual p value, m is the total number of tests, and Q is the false discovery rate, 0.05. Finally, we compared the original p values to the calculated q values. Attributes with a p value smaller than the corresponding q value and less than 0.05 were considered to have significant differences (which we highlighted in red in Tables 4 and 5).

To examine potential disparities across various groups of one algorithm, we employed a systematic approach. Initially, we categorized algorithm predictions based on their correctness, assigning a value of "1" to instances where the algorithm accurately predicted the ground truth and a value of "0" to instances where the algorithm falsely predicted the ground truth. Subsequently, we applied the Mann-Whitney test in conjunction with the Benjamini-Hochberg correction to different subsets of the "0" and "1" values to evaluate the following three hypotheses. First, we conducted a thorough analysis to determine whether the algorithms exhibited comparable accuracy in predicting the ground truth for both the protected and unprotected groups, aiming to evaluate potential disparities in accuracy. To achieve this, we performed the Mann-Whitney test with the Benjamini-Hochberg correction on the complete set of "0" and "1" values. Second, our assessment focused on whether the algorithms demonstrated similar false negative rates in predicting the

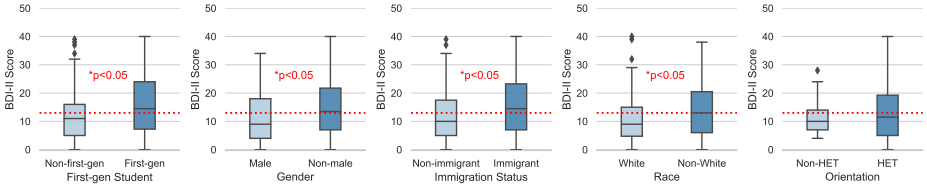
ground truth for both the protected and unprotected groups, to identify potential disparities in false negative rates. To accomplish this, we conducted the same test on the subset of "0" and "1" values where the ground truth labels were positive. Similarly, we proceeded to investigate whether the algorithms displayed comparable false positive rates for both the protected and unprotected groups. This was achieved by applying the same test on the subset of "0" and "1" values where the ground truth labels were negative.

A.3 Comparisons of Depression Scores for Different Groups of Four Datasets.

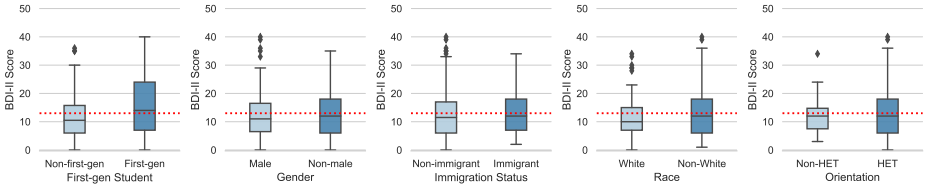
Fig. 4. Comparisons of depression (BDI-II) scores for different groups of four datasets. The red dotted line indicates the cutoff point (i.e., 13) for BDI-II scores, which is used to distinguish between students with at least mild depressive symptoms (BDI-II score >=13) and those without (BDI-II < 13). Significance levels after Benjamini-Hochberg (B-H) correction are marked with an asterisk ($^*p < 0.05$) in red on the subplot. First-gen, BA, and HET represent first-generation college students, bachelor, and heterosexual, respectively.



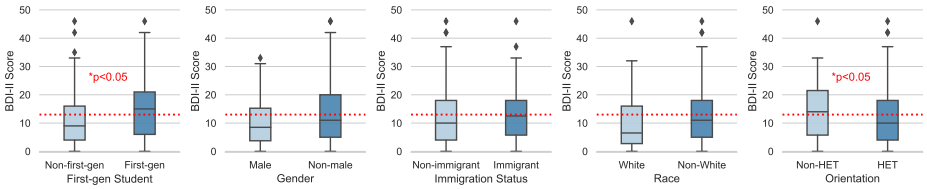
(a) DS1 (Year 2018, Pre-COVID)



(b) DS2 (Year 2019, Pre-COVID)



(c) DS3 (Year 2020, COVID Year)



(d) DS4 (Year 2021, COVID Year)

Table 9. Summary of bias changes with the addition of sensitive attributes in the training and testing process in terms of **disparity in accuracy**. This table provides an overview of bias alterations resulting from the inclusion of sensitive attributes during the training and testing processes, using disparity in accuracy as the fairness metrics. It encompasses bias amplification and reduction for each sensitive attribute across the four datasets. The comparison highlights the consequences of adding or excluding sensitive attributes in training and testing. Extra bias is denoted in red, while reduced bias is highlighted in green. For instance, considering the Xu_interpretable algorithm, Tables 4 and 5 present fairness evaluation outcomes before and after incorporating data related to first-generation college student status. When this sensitive attribute is introduced, an additional bias towards gender emerges in DS4, indicated by the label "1" in this table.

Algorithm	**Added Attributes**	**Sensitive Attributes**				
		First-gen college student	Gender	Immigration Status	Race	Sexual Orientation
Wahle *et al.*	First-gen college student	1	-1	0	0	0
	Gender	1	1	0	0	0
	Immigration Status	1	-1	0	0	0
	Race	0	-1	0	0	0
	Sexual Orientation	1	0	0	0	0
Saeb *et al.*	First-gen college student	-1	-1	-1	-1	-1
	Gender	-2	1	-1	-1	-1
	Immigration Status	0	0	-1	-1	0
	Race	0	0	0	0	0
	Sexual Orientation	-1	0	0	-1	-1
Farhan *et al.*	First-gen college student	0	0	0	0	0
	Gender	0	1	0	0	0
	Immigration Status	0	0	1	0	1
	Race	0	-1	0	1	1
	Sexual Orientation	0	1	0	0	0
Canzian *et al.*	First-gen college student	1	0	0	0	-1
	Gender	0	0	0	0	0
	Immigration Status	-1	-1	0	0	0
	Race	0	0	0	0	0
	Sexual Orientation	0	-1	2	0	0
Wang *et al.*	First-gen college student	1	0	-1	1	-1
	Gender	0	1	0	1	0
	Immigration Status	0	0	-1	1	0
	Race	0	1	-1	0	0
	Sexual Orientation	1	1	-1	0	0
Lu *et al.*	First-gen college student	0	1	0	2	1
	Gender	-1	0	0	1	0
	Immigration Status	0	0	0	1	0
	Race	0	0	0	1	0
	Sexual Orientation	-1	0	0	0	1
Xu_interpretable *et al.*	First-gen college student	0	1	0	0	0
	Gender	0	0	0	0	0
	Immigration Status	0	1	0	0	0
	Race	0	1	0	0	0
	Sexual Orientation	0	1	0	0	0
Xu_personalized *et al.*	First-gen college student	0	0	0	0	0
	Gender	0	0	0	0	0
	Immigration Status	0	0	0	0	0
	Race	0	0	0	0	0
	Sexual Orientation	0	0	0	0	0

Table 10. Summary of bias changes with the addition of sensitive attributes in the training and testing process in terms of **disparity in false negative rates**. This table provides an overview of bias alterations resulting from the inclusion of sensitive attributes during the training and testing processes, using disparity in false negative rates as the fairness metrics. It encompasses bias amplification and reduction for each sensitive attribute across the four datasets. The comparison highlights the consequences of adding or excluding sensitive attributes in training and testing. Extra bias is denoted in red, while reduced bias is highlighted in green.

Algorithm	**Added Attribute**	**Sensitive Attribute**				
		First-gen college student	Gender	Immigration Status	Race	Sexual Orientation
Wahle *et al.*	First-gen college student	2	0	0	1	0
	Gender	0	0	0	0	0
	Immigration Status	0	0	1	1	0
	Race	0	0	0	0	0
	Sexual Orientation	0	0	0	0	1
Saeb *et al.*	First-gen college student	3	2	0	2	-2
	Gender	1	2	0	1	-1
	Immigration Status	0	1	0	0	-1
	Race	0	0	0	2	-1
	Sexual Orientation	0	0	0	0	0
Farhan *et al.*	First-gen college student	1	0	-1	0	1
	Gender	0	0	-1	0	0
	Immigration Status	0	1	2	0	1
	Race	0	0	-1	0	0
	Sexual Orientation	0	0	0	0	1
Canzian *et al.*	First-gen college student	2	1	0	0	0
	Gender	0	0	0	0	0
	Immigration Status	0	0	3	1	1
	Race	1	0	0	1	0
	Sexual Orientation	0	0	0	0	1
Wang *et al.*	First-gen college student	1	0	0	1	0
	Gender	0	0	0	0	0
	Immigration Status	0	0	0	0	0
	Race	0	0	0	1	0
	Sexual Orientation	0	0	0	0	0
Lu *et al.*	First-gen college student	1	1	0	1	0
	Gender	0	2	0	0	0
	Immigration Status	-1	1	1	0	0
	Race	-1	0	0	3	0
	Sexual Orientation	-1	0	0	0	0
Xu_interpretable *et al.*	First-gen college student	-1	0	0	0	0
	Gender	0	1	0	0	0
	Immigration Status	-1	0	0	0	0
	Race	-1	0	0	0	0
	Sexual Orientation	-1	0	0	0	0
Xu_personalized *et al.*	First-gen college student	0	0	0	0	0
	Gender	0	0	0	0	0
	Immigration Status	0	0	0	0	0
	Race	0	0	0	0	0
	Sexual Orientation	0	0	0	0	0

Table 11. Summary of bias changes with the addition of sensitive attributes in the training and testing process in terms of **disparity in false positive rates**. This table provides an overview of bias alterations resulting from the inclusion of sensitive attributes during the training and testing processes, using disparity in false positive rates as the fairness metrics. It encompasses bias amplification and reduction for each sensitive attribute across the four datasets. The comparison highlights the consequences of adding or excluding sensitive attributes in training and testing. Extra bias is denoted in red, while reduced bias is highlighted in green.

Algorithm	**Added Attribute**	**Sensitive Attribute**				
		First-gen college student	Gender	Immigration Status	Race	Sexual Orientation
Wahle *et al.*	First-gen college student	3	0	1	1	0
	Gender	-1	1	0	0	0
	Immigration Status	0	0	1	0	0
	Race	1	0	0	2	0
	Sexual Orientation	0	0	0	1	1
Saeb *et al.*	First-gen college student	4	-1	0	1	0
	Gender	1	2	1	2	1
	Immigration Status	1	1	2	3	1
	Race	1	0	1	1	0
	Sexual Orientation	0	-1	0	0	1
Farhan *et al.*	First-gen college student	1	1	0	1	0
	Gender	0	2	0	1	-1
	Immigration Status	0	0	1	-1	-1
	Race	0	0	0	-1	-1
	Sexual Orientation	0	1	0	-1	-1
Canzian *et al.*	First-gen college student	1	-1	1	0	-1
	Gender	0	0	0	0	0
	Immigration Status	-1	-1	3	1	-1
	Race	0	0	0	-1	0
	Sexual Orientation	-1	-1	1	0	0
Wang *et al.*	First-gen college student	2	0	0	0	-1
	Gender	0	0	0	0	-1
	Immigration Status	0	0	1	0	0
	Race	1	1	0	1	-1
	Sexual Orientation	0	0	0	0	-1
Lu *et al.*	First-gen college student	3	1	-1	1	1
	Gender	0	3	-1	0	-1
	Immigration Status	1	0	1	1	-1
	Race	2	0	0	2	-1
	Sexual Orientation	0	0	0	0	0
Xu_interpretable *et al.*	First-gen college student	0	1	0	0	0
	Gender	0	0	0	0	0
	Immigration Status	1	1	0	0	
	Race	1	1	0	0	0
	Sexual Orientation	1	1	0	0	0
Xu_personalized *et al.*	First-gen college student	0	0	0	-1	0
	Gender	0	0	0	-1	0
	Immigration Status	0	0	0	-1	0
	Race	0	0	0	-1	0
	Sexual Orientation	0	0	0	-1	0