Conceptualizing Algorithmic Stigmatization

Nazanin Andalibi
andalibi@umich.edu
School of Information
University of Michigan
Ann Arbor, MI, USA

Cassidy Pyle
cpyle@umich.edu
School of Information
University of Michigan
Ann Arbor, MI, USA

Kristen Barta
kbarta@umich.edu
School of Information
University of Michigan
Ann Arbor, MI, USA

Lu Xian
xianl@umich.edu
School of Information
University of Michigan
Ann Arbor, MI, USA

Abigail Z. Jacobs
azjacobs@umich.edu
School of Information and Center for the Study of Complex Systems
University of Michigan
Ann Arbor, MI, USA

Mark S. Ackerman
ackerm@umich.edu
School of Information and Department of Computer Science and Engineering
University of Michigan
Ann Arbor, MI, USA

ABSTRACT

Algorithmic systems have infiltrated many aspects of our society, mundane to high-stakes, and can lead to algorithmic harms known as representational and allocative. In this paper, we consider what stigma theory illuminates about mechanisms leading to algorithmic harms in algorithmic assemblages. We apply the four stigma elements (*i.e.*, labeling, stereotyping, separation, status loss/discrimination) outlined in sociological stigma theories to algorithmic assemblages in two contexts : 1) "risk prediction" algorithms in higher education, and 2) suicidal expression and ideation detection on social media. We contribute the novel theoretical conceptualization of *algorithmic stigmatization* as a sociotechnical mechanism that leads to a unique kind of algorithmic harm: *algorithmic stigma*. Theorizing algorithmic stigmatization aids in identifying theoretically-driven points of intervention to mitigate and/or repair algorithmic stigma. While prior theorizations reveal how stigma governs socially and spatially, this work illustrates how stigma governs *sociotechnically*.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**;

KEYWORDS

stigma, algorithms, algorithmic decision-making, algorithmic harm, discrimination, labeling, stereotyping, separation, algorithmic stigma, representational harm, allocative harm, higher education, risk prediction, suicide ideation, mental health, social media, theory

ACM Reference Format:
Nazanin Andalibi, Cassidy Pyle, Kristen Barta, Lu Xian, Abigail Z. Jacobs, and Mark S. Ackerman. 2023. Conceptualizing Algorithmic Stigmatization. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3544548.3580970

CHI'23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3580970

1 INTRODUCTION

Algorithmic systems are increasingly pervasive in diverse contexts, ranging from mundane to high-stakes, and have implications for individuals, communities, organizations, and societies. These algorithmic systems have the potential to facilitate positive outcomes (*e.g.*, detecting and removing online hate speech, detecting and labeling potential misinformation) for impacted groups [135, 153, 188, 192]. Yet, they can also perpetuate and/or lead to algorithmic harms, including allocative harms [16] (*i.e.*, unjust distribution of resources and opportunity) and representational harms [16, 103] (*i.e.*, harmful representations along identity lines). More broadly, algorithms can perpetuate existing social inequities and create new ones [20, 21, 97, 141]. For example, the term "algorithms of oppression" [141] describes sexism and racism embedded in search engines' representation of Black women and girls. Algorithms underpinning social media news feeds can contribute to algorithmic symbolic annihilation [8] and harmful invisibility [32] of already marginalized groups. Facial recognition algorithms fail to recognize Black faces [140, 150] and associate negative emotions to Black faces regardless of a person smiling [155]. Search algorithms on employment websites can facilitate allocative harm by obscuring relevant work opportunities [65]. Algorithms in the child-welfare system cause harm to the very practice of social work [160]. Examples of harmful algorithmic systems are countless across domains.

The term algorithm can be elusive. Algorithms' definitions range from purely technical [189] to sociotechnical [79, 165]. In its most technical sense, an algorithm entails rules to perform a task based on some input data. Rather than taking the algorithm as one technological artifact detached from its social surroundings, we draw from Science and Technology Studies (STS) and critical data studies approaches and take the algorithm as part of a larger sociotechnical assemblage of both social and technical actors [79, 166]: the algorithmic assemblage [165]. The algorithmic assemblage entails human and non-human elements (and their encounters) in "sociomaterial entanglements whereby the algorithmic system is made and enabled to work in practice" [165] and is a valuable lens because

it affords a sociotechnical analysis of associated algorithmic artifacts. Considering both human and non-human elements allows us to better pinpoint algorithmic outcomes and mechanisms through which they come to be. In algorithmic assemblages, the algorithm itself is "but one element of the broader sociotechnical assemblage in which it is embedded" [165].

The sources of—and thus the locus of intervention for—algorithmic harms have often been framed as a technical challenge [59, 69, 175], however, there is growing substantial evidence establishing existing social injustices as sources for algorithmic harms [25, 86, 97, 156]. Still, the *mechanisms* leading to these harms and subsequent ways to interrupt, mitigate, and redress said harms remain under-theorized. Theorizing algorithmic harms and mechanisms through which they come to be, we argue, is an important step towards naming, redressing, and mitigating them. In this article, we take stigma theory from sociology [121] as a point of departure to theorize how some algorithmic harms are enacted and mediated by algorithmic assemblages [165] in a mechanism we define as *algorithmic stigmatization*.

Understood as a sociological mechanism, stigmatization generates social difference to enforce and reproduce inequality [144]. Scholars have defined stigma in myriad ways [80, 95, 121, 144, 181, 182]. Link and Phelan's synthesis [121] suggests that stigma exists when its elements (*i.e.*, labeling, stereotyping, othering/separating, status loss/discrimination) *co-occur* in the presence of social, political, and economic power [121]. Labeling involves distinguishing and labeling differences; stereotyping involves linking a label to negative values and stereotypes; and separation involves labels that connote or enforce people being kept "down, in and away" [120]. Essentially, "people are stigmatized when the fact that they are labeled, set apart, and linked to undesirable characteristics leads them to experience status loss and discrimination" [121].

In this article, we consider two algorithmic assemblages as diverse cases from which to explore algorithmic stigma: a) so-called "risk prediction" algorithms in higher education, and b) suicidal expression and ideation detection on social media.[1] In each case, we draw from the respective literature and other documentation to examine how stigma elements (*i.e.*, labeling, stereotyping, separating, discrimination/status loss) manifest in algorithmic assemblages. Our analysis shows how the four stigma elements manifest in these two cases, demonstrating the novel theoretical conceptualization of *algorithmic stigmatization* as a sociotechnical mechanism that leads to a unique algorithmic harm: *algorithmic stigma*.

We define algorithmic stigma as the type of stigma and algorithmic harm that is mediated, perpetuated, or sometimes created by/in algorithmic assemblages as a sociotechnical process; we therefore argue that stigma(tization) is *sociotechnical*, and not solely social as prior theories [80, 121, 147] suggest. The theoretical frame of algorithmic stigmatization reveals how allocative and representational harms contribute to and are entangled within algorithmic stigma, but that algorithmic stigma is distinct in that it occurs when *all* four stigma elements converge. While representational and allocational harms as concepts are useful abstractions that can describe many different harms, we argue that these abstractions, by themselves, are inadequate to fully describe and analyze the harms caused by stigma and stigmatization as manifested in algorithmic assemblages. Indeed, stigma is complex; many representational and allocational harms can accrue from stigmatization and stigma (see figure 1). As such, we assert that conceptualizing and recognizing algorithmic stigma(tization) separately from other harms with allocational and representational dimensions is a useful contribution to scholarship on algorithms' societal and ethical implications.

We argue that our theorizing of algorithmic stigmatization aids in identifying distinct, theoretically-driven points of intervention to mitigate and/or repair algorithmic stigma. We discuss how reparative approaches [60] may afford beginning to disrupt algorithmic stigmatization. Ultimately, theorizing algorithmic stigmatization contributes to broader sociological understandings [182] of how stigma reifies inequality. By articulating the mechanism through which algorithmic stigma emerges (*i.e.*, algorithmic stigmatization), this work reveals *how* stigma governs *sociotechnically*, extending previous theoretical insights suggesting how stigma governs socially and spatially [147].

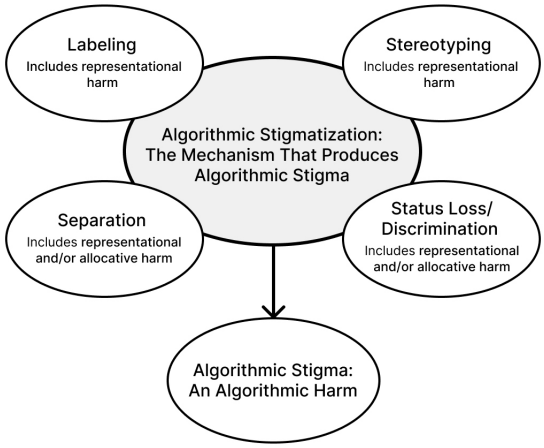

Figure 1: This figure shows how the *algorithmic stigmatization* mechanism produces *algorithmic stigma*. Algorithmic stigma(tization) is mediated by/in algorithmic assemblages. Algorithmic stigmatization is the convergence mechanism of labeling, stereotyping, separation, and status loss/discrimination in an algorithmic assemblage, leading to a distinct algorithmic harm: algorithmic stigma. Along the way, the diagram also depicts where previously known algorithmic harms (*i.e.*, representational, allocative) are relevant within the algorithmic stigmatization process, but that algorithmic stigma is distinct in that it is produced when all four stigma components are present.

2 RELATED WORK

2.1 Theories and Conceptualizations of Stigma

Stigma has been a topic of interest in disciplines ranging from sociology [121, 147, 180, 182], public health [48], social psychology

[1] These cases are illustrative in highlighting algorithmic stigma and stigmatization as novel theoretical concepts, but they are not necessarily exhaustive. Just as elements of stigma manifest differently across these cases, we expect that they may manifest differently in other algorithmic assemblages. Future work could systematically examine algorithmic stigmatization's patterns across diverse cases.

[96], law [34], and computing [6, 37, 39, 124], among others. Goffman's book, *Stigma: Notes on the Management of Spoiled Identity*, popularized a definition of stigma as "an attribute that is deeply discrediting" [80, p. 3]. Since then, social psychological perspectives of stigmatization traditionally focused on how people mentally construct categories and link categories to stereotypes [96]. Goffman's conceptualization of stigma—later critiqued for neglecting the role of social structure [144, 161] and for being apolitical and sometimes offensive [180]—proved to be foundational for an enduring legacy of research on the process and consequences of stigmatization. Modern sociological perspectives on stigmatization have moved this legacy forward while centering the complex dynamics of stigma, stigmatization, and the way power is enacted [182].

Sociological understandings of stigma address the impact of power and social structures to enforce social differences, beyond (and in addition to) intrapersonal and interpersonal dynamics. Invoking a sociological perspective, Link and Phelan's [121] conceptualization of stigma situates stigmatization as a *social process* by distilling the process of stigmatization into four primary, co-occurring components through which power is enacted: labeling, stereotyping, separation, and discrimination/status loss. In doing so, they provide a consistent definition, grounded in the lived experiences of the stigmatized, that attend to meso- and macro-level, in addition to micro-level, interactions.

Beyond Link and Phelan's [121] contribution to stigma literature, myriad scholars have advanced theoretical understandings of stigmatization as a social process embedded in larger social and power structures. For instance, in the context of HIV/AIDS stigma, Parker & Aggleton [144] highlight how stigma draws from and exacerbates racial, gendered, and classed inequalities to argue that stigma interventions must similarly take a structural perspective. Similarly, Scambler [161] focuses on epilepsy and HIV stigma to note how stigmatized attributes are embedded in cultural norms and social structures of class, gender, and ethnicity. Subsequent sociological accounts of stigma have incorporated Foucault's notion of governmentality [94] to describe stigma's individual and collective impacts. Conceptual framings such as 'stigma power' [120] have emerged to describe stigma processes as indirect and deeply hidden and embedded in taken-for-granted social and cultural circumstances. Mre recently, scholars such as Tyler have proposed the notion of stigma as a governmental technology used to further dehumanize the stigmatized [180] and as a cultural and political economy [182]. Because of the way these sociological accounts of stigma engage meaningfully with power, identity, and social structure in addition to interpersonal and sometimes intrapersonal dynamics, we posit that sociological conceptualizations of stigma are particularly useful in thinking about algorithms' implications.

Link and Phelan [121] argue that stigmatization involves four elements that must all exist for stigmatization to occur, though they need not occur in order. Importantly, these elements are not entirely distinct in practice; they are separable only analytically. Yet, distinguishing these elements is useful for identifying and analyzing how they may manifest in diverse contexts. We describe stigmatization's four elements next.

Distinguishing and Labeling Difference. Link and Phelan [121] argue that the process of distinguishing and labeling difference is a social process that involves "substantial oversimplification" (p. 367) of individual characteristics and attributes into a "label." Labels attached to behaviors or individuals construct social deviance, *i.e.*, deviance from the "mainstream" [18]. In contrast to Goffman [80], who uses the term "attribute," the term "label" hints at the ways that difference is *ascribed* to someone in a social process rather than being inherent to an individual. Labeling is a social sorting process done to someone, and while the label a person receives may be accurate or consistent with one's perception of oneself, often it is not. Yet the act of labeling enacts power by ascribing those attributes what Hacking describes as "making up people" [90].

Stereotyping - Associating Difference with Negative Attributes. Beyond demarcating "difference," stigmatizing labels are linked to socially situated stereotypes. As Link and Phelan [121] note, the label performs the work of "linking a person to a set of undesirable characteristics that form the stereotype" (p. 369). This social psychological process is where labels take on connotative meanings. For example, when considering the stigmatization faced by people with schizophrenia, Angermeyer & Matschinger [11] note that the label of schizophrenia "increases the likelihood that someone suffering from the disorder is considered as being unpredictable and dangerous" (p. 394) based on the link between the label (schizophrenia) and stereotypes (danger, volatility).

Separation. Linking labels to negatively-valenced stereotypes facilitates separation of the stigmatized group. This separation separates in-group from out-group, "us" from "them" [121], keeping the stigmatized group "down, in and away" [120]. Thus the separation of groups can enforce structural discrimination between groups [12]: Stereotyping allows one to see people as inhabiting group-level characteristics rather than understanding their full-fledged humanity, which makes it easy to group people into an out-group that can be kept physically and/or socially separated —which can contribute to structural discrimination [12]. To return to the example of people with schizophrenia, people who more strongly associate the label of schizophrenia with stereotypes of danger and volatility more strongly prefer to distance themselves socially from people said to be "schizophrenic" [11, 12].

Status Loss and Discrimination. Labeling, stereotyping, and separation result in negative consequences for the stigmatized person(s). Status loss refers to "a general downward placement of a person in a status hierarchy" [121], bringing with it a lack of opportunities. Alongside status loss, other individual consequences of stigmatization include reduced self-esteem, lack of a clarified identity, and feelings of unworthiness [46]. These consequences align with the concept of self-stigma, or the internalization of stigma that results in diminished self-esteem and self-efficacy [190]. Returning to the example of stigma surrounding schizophrenia, status loss and discrimination may occur when people with schizophrenia grapple with over-protection or infantilization [82] and social isolation due to the social distancing described above [11].

Stigmatization and status loss carry implications for one's access to resources. Discrimination exists at individual, interpersonal, and structural levels. The individual level of discrimination dominates stigma literature from social psychology, emphasizing interpersonal interactions in which "Person A" discriminates against "Person B." Sociologists such as Link and Phelan [121] draw from discussions of institutional racism to expand the discriminatory component of stigmatization in a way that includes structural features of one's

environment (see also [122, 182]). They provide the example of ableist work environments that preclude disabled people from being able to work as one example of structural discrimination [121]. Subsequent work has identified how stigmatization leads to meso- and macro-level outcomes ranging from mental and physical health disparities to economic inequality [46]. Thus, discrimination as one component of stigmatization can exist irrespective of whether individual or interpersonal actions are discriminatory.

Scholars have leveraged the four components of the stigmatization process revealed by Link & Phelan [121] to examine stigmatization experienced by diverse groups (*e.g.*, LGBTQ+ populations [133], people with mental health conditions [52, 53], and people impacted by the criminal justice system [142]). While stigma has been a topic in computing, for example examining how stigmatized groups use or do not use technology [6, 9, 91, 124, 193], less of this discourse has systematically examined the stigmatization *process* itself, nor has it deeply interrogated the sociotechnical properties of technology (*e.g.*, algorithmic assemblages) that may facilitate the stigmatization process. Of the papers that cite Link & Phelan's [121] canonical work, a small fraction contends with the role of technology in shaping the stigma process, with stigma theory and algorithms discourse, in particular, remaining mostly disjoint [2].

Overall, the literature on stigma establishes stigmatization as a social process that creates demeaning differences—entailing interpersonal and structural aspects reviewed above. But stigma also acts *spatially*, such that stigma associated with place further *stigmatizes*. Thus place impacts individuals' sense of self, interpersonal and inter-regional dynamics, as well as individual and community-level outcomes such as health [63, 171]. For instance, during the Great Recession, Detroit residents were impacted by the 'symbolic degradation associated with their city' [84]. Spatial stigma not only impacts individuals and communities, but also research and policy: for example, spatial stigma is under-researched and leads to less and less useful public health interventions for stigmatized communities [92]. Wacquant and others trace how space can be used to create a social distinction that is then used for systematic disinvestment and/or increased punitive control by the state [182, 186]. Thus the way that space can create social distinction [185, 186] is indeed an example of stigma power: "keeping [stigmatized] people in, down and away" [120].

The literature reviewed here establishes stigma as a social structural process that generates and reinforces inequities, and a process that acts socially and spatially. Building on this legacy of stigma scholarship, in this paper we will theorize how stigma is enacted *sociotechnically*, motivating our consideration of stigmatization as a theoretically rich concept with the potential to explain the mechanisms leading to stigma-related harms in algorithmic assemblages.

2.2 Algorithmic Harms

Mitigating algorithmic harms is impossible without naming said harms, identifying their sources, recognizing their implications, and articulating the mechanisms leading to them. There is a growing body of research across fields including Human-Computer Interaction (HCI), Computer-Supported Cooperative Work (CSCW), Fairness, Accountability, and Transparency (FAccT), and critical algorithms studies concerned with harms posed by algorithmic systems in diverse contexts, approaches to addressing these harms, and critiques thereof.

Of note, in 2017, Barocas et al. [16] and Crawford [105] introduced a framework for thinking about how the negative consequences of automated systems manifest as allocational and representational harms. [3] Broadly, allocational harms arise when an automated system allocates resources (*e.g.*, money, credit) or opportunities (*e.g.*, jobs) unfairly to different social groups, or when opportunities or resources are withheld from certain groups. For instance, allocational harms exist when people are prevented from receiving resources such as income in the case of gig work algorithms (de)prioritizing certain workers in allocating gig opportunities [137]. Similarly, those wishing to receive a ride via a rideshare service may be (de)prioritized based on factors ranging from identity to technical components (*e.g.*, how charged their phones are), and may face algorithmically-driven "surge pricing" [137]. Allocational harms can also be less obvious. For instance, business rating platform algorithms may unfairly deprioritize certain businesses in searches or prioritize businesses that receive poor ratings over those that receive higher ratings [66]. As in the case with business rating algorithms, these allocational harms (*i.e.*, ratings or exposure to customers) can often translate into financial harms, as businesses struggle to recruit and maintain customers who are influenced by biased rating algorithms [66].

Representational harms manifest when a system (*e.g.*, a search engine, image captioning system, social media news feed) misrepresents, excludes, suppresses, or demeans social groups [103, 107, 169, 187]. In other words, algorithmic representational harm [103] is harm experienced when being subjected to *algorithmic symbolic annihilation* [8], which Andalibi and Garcia define as "how algorithms perpetuate normative and stereotypical narratives about phenomena, where what they account for has power and authority, and what they do not account for does not." [8] In the context of social media news feed algorithms, Karizat et al. [103] define *algorithmic representational harm* as the kind of harm that "users experience as a result of being rendered invisible, trivialized, suppressed, or otherwise further marginalized on the basis of their identities and the algorithm's understanding of their identities."

In the context of computer vision algorithms, Katzman et al. [107] decompose representational harms into categories including denying people the opportunity to self-identify and erasing or demeaning social groups. Automated Gender Recognition (AGR) technology (*i.e.*, technology that "purports to allow the automatic, computational identification of a person's gender from photographs or videos") [110] is, for instance, well suited to producing and exacerbating representational harm in the forms that Katzman et al. [107] explore. As Keyes [110] notes, AGR's treatment of gender as binary and physiological facilitates the normative, harmful erasure of trans and non-binary people – which we interpret as a type of algorithmic symbolic annihilation [8], leading to representational

[2]We used Google Scholar's "Cited By" feature to find work that cites Link and Phelan's canonical 2001 paper and also uses the term "technology" in the paper. In doing so, we discovered that these works primarily consider assistive technologies vis-a-vis stigmatization.

[3]In our case analyses, we refer to these harms where relevant.

harms. What these definitions of representational harm have in common is the notion that how humans are accounted for and representation (and lack thereof) matter in examining algorithmic systems' implications.

We note that allocational and representational harms are not independent and are usually accompanied by each other. Consider, for example, a search engine disproportionately displaying advertisements about criminal records when common African American names are searched [176]. If this leads to racial discrimination against loan applicants, that would constitute an allocational consequence. Even if it does not, the perpetuation of racial biases still entails representational harm [21, 176]. Additionally, in the case of AGR, representational harm surfaces in the erasure of non-binary gender and the problematic notions of biological essentialism that undergird these algorithms [162]. Yet, when deployed, for instance, in (gendered) bathrooms to police who goes in and out in the auspices of "safety", trans and non-binary people can experience discrimination, policing, harm, and violence at the hands of other bathroom patrons as well as the police [110]. Another example is targeted advertising, which may involve oversimplified algorithmic relevancy models that display weight loss advertisements on the social media feeds of those who have searched "intuitive eating" content (which is expressly anti-diet culture), reducing the visibility of content that aims to chip away at diet culture in favor of content that reifies it [74].

Purely technical fixes to address algorithmic harms fail to capture the sociotechnical context of the sources of harms, and therefore fail to intervene effectively. Technical interventions include further excluding proxies for protected variables [77] and constructing corrections to disparities against protected groups created by input variables [152]. Yet such interventions offer no guarantee: although protected categories may be excluded from algorithmic inputs, their proxies may still remain and be fed into algorithms [17]. In fact, the analysis of allocational harms that merely focuses on outcomes that protected groups obtain elides the discussion of what disadvantages socially constitute those protected categories in the first place [99, 104]. That is, the sources of harms experienced by protected (and other marginalized) groups are rooted in social relations and practices, and naive, purely technical, interventions fail to capture harms' complex sources.

More substantive approaches to algorithmic harms attend to social context more broadly, drawing from STS [64, 132], critical race theory [21, 93], feminist scholarship [8, 106, 162, 191], legal scholarship [137, 178], and political theory [49, 177]. For instance, Hoffmann [97] highlights the need to address the structural conditions that algorithmic systems help reify and reproduce. Since it is the structural conditions that produce individuals with different positions of advantages and disadvantages, Green [86] calls for evaluating fairness by taking into account such structural conditions reflected by data to better understand the implications of algorithmic decisions. In a similar vein, scholars such as Eubanks [68], Gangadharan [76], and Zeide [195] bring to light the systematic curation of opportunities, leading to exclusion, created by automated assessment algorithms beyond specific decision points. Emphasizing representational harms, Katzman et al. [107] and Wang et al. [187] highlight the challenges inherent in measuring substantive, representational harms such that they can be mitigated at all.

In summary, addressing algorithmic harms in a meaningful way requires understanding the sources and implications of algorithmic harms and mechanisms leading to them. Such an understanding demands the substantive consideration of the interactions between algorithms and social contexts where algorithms are embedded, as harms are context-dependent, impacting individuals and communities in particular ways based on circumstances [132]. Our case analyses draw on the concept of algorithmic assemblages [165] to account for the *sociotechnical* components that facilitate algorithmic harm. Scholars have proposed several lenses to enable this sort of investigation, including intersectionality rooted in Black feminism and reparation [47, 58, 60], slow violence [74, 139], co-production [103], justice [97], value-sensitive design [71, 72, 196], and critical disability [22] approaches among others. This paper examines what stigma theory contributes to this discourse surrounding algorithmic harms and addressing them. While representational and allocational harms, as concepts, are useful abstractions that can describe many different harms, we argue (and will show in our case analyses) that these abstractions, by themselves, are inadequate to fully describe and analyze the harms caused by stigma and stigmatization as manifested in algorithmic assemblages. As we will show in our case analyses, stigma theory illuminates novel insights about algorithmic harms and mechanisms leading to stigma. These insights are important, as to begin to mitigate algorithmic harms, it is important to name, describe, and articulate how those harms emerge – this paper's focus.

3 CASE ANALYSES: STIGMA AND ALGORITHMIC ASSEMBLAGES

In this section, we draw from prior work and existing documentation to present two case analyses that demonstrate the utility of the stigma lens in understanding algorithmic harms and how they come to be. We first introduce each case, then describe how the four stigma elements (*i.e.*, labeling, stereotyping, separation, status loss/discrimination) [121] co-occur in algorithmic assemblages. We note that while these elements form the analytical categories for our analyses, they are inseparable and intertwined in practice. In discussing each of the four elements for each of the two cases, we also note where other algorithmic harms manifest along the way.

We chose these two cases because they encompass two arguably diverse cases in which algorithmic harms may be relevant, and on which there is substantial research and documentation to draw from for our analysis. These cases represent a range in how various algorithmic harms may manifest. While the ultimate outcome of risk prediction algorithms in higher education may be access to opportunity felt in a very tangible way (in addition to other harms, as we show), the manifestation of harms in the social media context are somewhat more obscured but still highly influential, such as by shaping access to social support and visibility. The institutions governing these assemblages are also different; universities in the first and social media platforms in the second. Lastly, how and the extent to which algorithms' use is evident to and experienced by those impacted by them varies across these cases, with perhaps more opaqueness in the social media case.

3.1 The Case of "Risk Prediction" Algorithms in Higher Education

"Risk prediction" algorithms in higher education aim to predict a student's risk of failing and/or dropping out of a college or university. The emphasis on "risk" in education, particularly higher education, emerged most strongly in response to the 1983 report "A Nation at Risk: The Imperative for Educational Reform" by the United States National Commission on Excellence in Education [31, 151]. The risk framing that figured prominently in this report pointed to the urgency of educational reform and tied the nation's economic stability to educational outcomes, particularly with respect to higher education and preparing the next generation of the American workforce. As such, student groups that were previously labeled as "culturally deprived" or "educationally disadvantaged" (*e.g.*, low-income students, minorities, student parents, etc.) began to don the new moniker of "at-risk" [151]. In line with this linguistic shift, in 1994 the Office of Educational Research and Improvement at the U.S. Department of Education created the National Institute on the Education of At-Risk Students (NIEARS) [31].

As a result, higher education institutions have taken it upon themselves to calculate and predict the level of "risk" a student is perceived to embody concerning potentially failing or dropping out of the institution. Brown [31] highlights the five activities in identifying and intervening in the life of the "at-risk" student: "1) identifying the specific populations who experience some negative outcome more than other populations of people; 2) isolating, in the population that experiences the negative outcome, the specific actors associated with the occurrence of that negative outcome; 3) categorizing those populations deemed more likely than other populations to experience the negative outcome as 'at-risk'; 4) designing and implementing interventions to eliminate, or at least buffer, the possible effects of the risk factors; and 5) evaluating how effective the risk intervention was in countering/buffering the effects of risk factors".

To identify specific populations of "at-risk" students, higher education institutions initially deployed early risk prediction models that relied on simple, fixed factors such as high school GPA, socioeconomic status, and SAT scores to forecast a single prediction within a particular time frame [115]. Although these models were able to predict retention with reasonable precision, they did so only at a single, fixed point in time. That is to say, these models were unable to incorporate fine-grained, shifting information when predicting student "risk" (operationalized as retention) [115].

In recent years, sophisticated AI/ML techniques have enabled contemporary risk prediction technologies to predict "risk" based on multiple, fluid data sources and types such as demographic information and academic records, facilitating real-time predictions that can help to facilitate more timely interventions on the part of the college or university [115]. The rise of virtual learning environments (VLEs) and Learning Management Systems (LMS) such as Canvas allow for the harvesting of even more data on how often a student logs in, views documents, and views discussion forums, which is then used to make predictions about "risk," often in conjunction with a student's demographic and academic records [38, 42]. Across all of these systems are upstream decisions about what constitutes "risk" and how student performance and risk should be operationalized.

3.1.1 Distinguishing and Labeling Difference: Institutional, Behavioral, and Demographic Data Mining. The first task of "risk" prediction algorithm deployment is problem formulation and gathering data to make inferences [146]. Data gathering often involves data mining, wherein large datasets are derived and sometimes combined to produce insights. Algorithmic systems have been critiqued for their "black box" [145] nature, or for the lack of transparency about how input data are collected and used. Labeling is fundamental to both developing algorithms and to what algorithms do in practice. In practice, the labeling process can appear in various stages of the development and use of algorithms: in problem formulation, where the task of "risk prediction" is clarified and operationalized, and in the data collection and algorithm development phase, where humans may annotate student data, infer labels of "risk," or sort data into buckets along some "risk" dimension [101, 146]. The decisions shaping labeling may be informed by the developer and/or data annotators' understanding of the world [61, 148].

In the case of higher education risk prediction algorithms, developers may imbue their algorithms with assumptions about higher education and college students that may be problematic. For instance, the uninformed idea that underrepresented students are *inherently* more at-risk of facing academic difficulties or of dropping out may lead developers to associate being a student of color, a first-generation student, and/or a low-income student with greater "risk". Yet, disparities in retention and academic difficulty can be attributed to external and structural factors like racialized wealth inequality [154] and hailing from underfunded schools and neighborhoods [194]. In fact, many algorithms of this sort consider demographic factors as risk indicators (*e.g.*, [5, 23, 114]).

More sophisticated machine learning approaches to risk prediction tasks have also triggered concerns around the development of training data sets (*e.g.*, [41, 61, 100, 148]). Training datasets require data annotators to assign annotations (*i.e.*, labels) to the data. These "labeled" datasets are then used to train machine learning models for prediction. Prior work (*e.g.*, [26, 148]) has explored challenges related to training data and processes, especially with data concerning people. For example, it has shown how training sets can be biased along dimensions such as race, gender, disability, nationality, etc. [73, 117] and that annotators' work is informed by values and priorities imposed on them by actors above "their stations" [134]; these priorities may include profit, standardization, and opacity [109]. Moreover, using these datasets have reifying effects [15, 111, 162].

In the higher education context, the data that serve as inputs for risk prediction algorithms are often *institutional data* derived from various systems and units across campus. Institutional data can include financial records, students' academic records, and records on how often students accessed services such as academic advising. Within the broader category of institutional data, *historical behavioral* data (*i.e.*, data on what a student has previously done) such as students' assignment submission records are commonly used (cite). *Non-academic* historical and behavioral data, such as how students use their ID card or library card to access buildings, check out library books, etc., is also sometimes collected and used

[108]. Demographic data such as gender, race, ethnicity, and socioeconomic status are also commonly used in conjunction with the factors mentioned previously. Specifically, identities that have been marginalized and underrepresented in academia (*e.g.*, students of color, first-generation students, low-income students) come to symbolize greater degrees of "risk" [85].

It is important to note that scholars have critiqued the use of historical, demographic, and behavioral data in predictive algorithms. They argue that using historical data to predict future behavior traps students in a loop regarding past behaviors [170], a core challenge in education (*e.g.*, [131]). Additionally, they note that the use of identity or demographic data often renders students' identities singular and static [170] and synonymizes marginalized identities with risk and deviance [85]. Finally, scholars note that (over)reliance on behavioral data privileges "behaviorism," an educational philosophy that has earned critique for configuring students as passive recipients of knowledge and assessing the degree of student effort and learning based on observable behaviors [168, 170].

Nevertheless, deploying risk prediction algorithms in higher education takes various forms and sources of data and introduces what Link & Phelan [121] would characterize as labels. Nuanced lived experiences are reduced to categorical identity or demographic variables (*e.g.*, Black, low-income). Students' learning experiences are reduced to observable grades on a transcript and devoid of context. The number of times an individual visits the library marks them as a high or low user, disregarding why this might be the case. Returning to Link & Phelan [121], such distinguishing and labeling of difference in the higher education context constitutes an oversimplification of the myriad granular differences among students and the stripping away of contextual factors that may have shaped these categorical difference markers.

3.1.2 Stereotyping: Risk Prediction as Deficit Framing. The second element of the stigmatization process outlined by [121] is stereotyping. In algorithmic assemblages, stereotyping can occur in several ways. First, people who interact with algorithm systems may engage in acts of stereotyping others as mediated by algorithms and can also develop understandings of stereotypes. For example, people who search the term 'criminal' in a search engine that algorithmically generates only images of Black faces may develop stereotypical understandings of race. Second, algorithm developers may encode stereotypes in algorithms. Indeed, it is well established that developers and data set annotators encode their values and understandings of the world into the data used to train AI/ML algorithms [61, 100, 148]. Third, algorithms may perpetuate stereotypes, such as the Google Image search algorithm's perpetuation of the stereotype of Black women as hypersexual [141].

In the context of higher education, the design and deployment of "risk prediction" algorithms contributes to stereotyping. That is, this process of associating labels (of risk) to negative values (as deficit) comprises both negatively-valenced inputs and the use and interpretation of the label "risk." Emphasizing "risk" is a form of deficit framing [149] that shifts responsibility for student success away from the environment and toward the individual student, with the potential to pathologize individual students who are deemed "at risk." Gray [85] notes that "since the early 90s, the term 'risk' has been widely deployed in arenas of education, health, and social services. Because it generally denotes a description of what is taking place in the lives of particular groups of people, risk analyses locate the problem in the students themselves or in their families or their histories". Regarding the university faculty and staff who interact with students, Rozycki [158] argues that the "at risk" label indicates a possible confrontation with something "undesirable."

Once student data is input into a "risk prediction" algorithm, the algorithm can then draw from this data and use predictive risk modeling to elicit a "risk level" [164]. Implicit or explicit assumptions about "at-risk" students affect how these risk levels are interpreted (negatively) by higher education personnel, as well as how they influence future interventions (*e.g.*, allocating attention or opportunities). Algorithmic stereotyping involves upstream, technical processes, including what kinds of data were collected and how risk is operationalized, and downstream, interpretive processes, such as how students who receive the label interpret it and what universities decide based on this label.

3.1.3 Separation: Implementing Interventions. In the case of higher education "risk prediction" algorithms, an algorithm might deem a college student "at risk," prompting higher education personnel to institute some form of intervention that performs the work of separation. These interventions can potentially separate so-called "risky" students from their less "risky" peers physically, mentally, and socially. As such, they can facilitate both representational and allocative harms.

Physically and mentally, universities can segregate "risky" students into remedial or "bridging" classrooms. Often, these remedial or bridging courses can hinder students' progress toward a degree and force them to engage in tedious rote memorization tasks instead of developing their critical thinking skills [4, 128, 129]. Separation of "at-risk" students into remedial classrooms may have allocative consequences when it hinders the opportunity of "at-risk" students to obtain a quality education and enact agency over their educational choices. If the university concludes that an "at-risk" student will not be successful, the university will suggest that the student switch to less "demanding" majors, avoiding putting resources into that student [24, 164]. In this way, the university allocates educational resources unfairly across "at-risk" students and their less "risky" counterparts. Separation also has psychological effects in discouraging a student from continuing with study when it physically (and thus socially) isolates the student from their campus peers [164]. With less social support and a potentially lowered sense of belonging, the isolated student may be less likely to pursue more technical majors than students who are not deemed "risky", as evidenced by prior work (e.g., [19, 29, 88, 174]). This may lead to disparities in the choice of major between these populations and downstream, higher-level effects in the demographic composition of students in technical majors.

Interventions can also *socially* separate "at risk" students from their less "risky" counterparts. Often, though not always and not a prerequisite for social separation, students are notified of their own "at risk" status through phone, email, or an online Dashboard [115]. For example, Lu et al. [123] describe an intervention in which teachers notify relevant students of their risk status and arrange for face-to-face consultations if deemed necessary. Similarly, the Course Signals (CS) program provides individual risk assessment

reports for students to peruse [13], and Student Explorer uses traffic light colors of red, yellow, and green to visualize risk levels for both teachers and students [112]. Once notified, students are left to think about how and why they represent risk in comparison to their less "risky" counterparts. When a student who embodies marginalized identities learns they are considered "at-risk," they may internalize the stereotype linking these marginalized identities to "risk" and "deviance," which can facilitate representational harm.

Interventions such as those reviewed above alert students to the fact that they have been classified as "at-risk." Importantly, for individuals, negative psychological effects from interventions can also occur when a student is *not* notified about their "at-risk" status. For example, when students are required to (*i.e.*, not given a choice) take remedial courses or to attend extra advising sessions while others are not, engaging in social comparison would also prompt them consider why they are required to do so while others are not. Additionally, requiring students to take extra coursework or attend advising meetings compromises the agency students can enact over their educational pathways, which may constitute an autonomy harm [45] (*i.e.*, a kind of representational harm) [16, 103, 105].

3.1.4 Status Loss and Discrimination: Interpersonal Discrimination, Psychological Harm, and Surveillance. Allegedly "at-risk" college students face status loss and discrimination. Beyond being identified and physically, mentally, and/or socially separated from their less "risky" peers, they can face various forms of status loss and discrimination. Interpersonal forms of status loss and discrimination operate via differential treatment by faculty, staff, and/or peers. Structural forms of status loss and discrimination operate via interventions enacted by the university that may surveil and police them, compromise their agency, and potentially impact the way they think about themselves (*e.g.*, internalized stigma [183, 190]) and behave as learners.

Notifying students, faculty, and staff about specific students' risk level can open possibilities for interpersonal discrimination between peers and between an "at-risk" student and university faculty and staff. Masood [126] notes how students "understand each other in terms of... subjectivities and tend to perform in ways that reproduce their subject positions". In the case of so-called "at-risk" students, this can mean that students affixed with the "at-risk" label can internalize stigma [183, 190], which can later shape the interactions they have with their peers who are deemed less "risky." This occurrence can create a problematic feedback loop wherein students are notified that they are "risky," internalize this idea, self-segregate, or put themselves down in interactions with others, which then opens them up to more opportunities to experience stigmatization they can, again, internalize.

Interventions that center students' "risk level" can create internal, psychological difficulties for students deemed risky, which constitutes allocative harm. In an empirical study, Perez [149] found that the frequency of receiving the negative "at-risk" label was significantly associated with students' negative academic self-perceptions, sense of belonging, and negative affect. Lawson et al. [116] also describe how emails sent to students informing them of their "at-risk" status could demotivate students by noting that a student is "more than likely going to fail" or "if you don't attempt this assessment, you will fail" (p. 961, 964), echoing Kruse & Pongsajapan's [113] assertion that interventions based on risk prediction can lead to "inefficiency, resentment, and demotivation" (p. 3). Interventions that communicate the negative descriptor of being "at-risk" to students may also be detrimental to their motivation and persistence in college. Such interventions which cause representational harm which can then facilitate undesirable allocative consequences by inadvertently exacerbating the disparities in persistence rates that interventions aim to ameliorate [149]. Hence, when students internalize the belief that they are "at-risk," and thus less likely to achieve academic success, these negative psychological effects may facilitate or exacerbate demotivation, which compromises their academic success and facilitates the very negative outcomes (*e.g.*, failing a course, dropping out of university) that predictive risk algorithms claim to want to prevent.

Additionally, interventions that contribute to and reinforce structural forms of discrimination often involve some degree of tracking and documentation, which can be conceived of as surveillance in a broader sense. Some scholars argue that the interventions that schools enact in response to a student's "at-risk" designation "profiles and pre-emptively punishes" them via "techniques of discipline and in the guise of remediation" [126]. Students placed in remedial classes or bridge programs as part of a kind of "at-risk" intervention that Masood [126] calls "dividing practices" are thus prescribed activities or repeated exercises that can serve as "techniques of control" that encourage self-discipline and self-regulation [126] and therefore compromise students' ability to enact agency and autonomy over their university experiences. This autonomy harm [45] is a subset of broader representational harms [16, 103]. Moreover, "risky" students' integration into remedial or bridge classes may restrict other educational opportunities [126], thus limiting their pathways to academic success and potentially contributing to disparities in educational outcomes between "at-risk" students and their "non-risky" counterparts. Furthermore, the control over students' behavior calls into question the agency they can enact over their education. Students' subsequent academic behaviors after being affixed with the "at-risk" label are closely monitored, tracked, and documented, constituting a problematic form of surveillance and policing [85, 126].

In sum, we have shown how the four stigma elements manifest in "risk" prediction algorithmic assemblages in higher education.

3.2 The Case of Suicidal Expression and Ideation Detection on Social Media

Numerous social media platforms have designed tools to identify content expressing mental health states, such as suicidal ideation. This identification triggers responses from the platform, such as offering resources (*e.g.*, chatline, hotline number) and moderating, removing, or deprioritizing content to maintain "community safety." The sophistication and automation of such tools vary and may include a combination of algorithm- and human moderator-driven methods. In this case study, we consider tools used on Meta's platforms, Facebook and Instagram, and identify how such tools facilitate algorithmic stigmatization.

3.2.1 Labeling Suicidal Ideation: Moderation and Classification. Human- and algorithmic content moderation classifies social media content and accounts by flagging content for removal or accounts

as sensitive. To come to classification decisions "involves inducing generalizations about features of many examples from a given category into which unknown examples may be classified" [83]. In other words, this classification process sorts content and accounts into different buckets—that is, this process *labels*. Labeling involves both automated and human-directed processes, as labels are socially determined [27] and may be applied through automated processes. Both social and automated aspects of algorithmic assemblages can contribute to harm, as we show throughout these case studies. Labeling is generally opaque to users and the public [30], such that there is little visibility afforded to users as to what the labels are (if they do indeed exist), how they are determined, and how they are algorithmically applied. Social media community guidelines may offer insight into generic labels of interest to platforms (*e.g.*, classifying content as containing nudity, hate speech, violence, etc.).

In the mental health context, content containing indications of suicidal ideation/intent or self-harm may be labeled as violent or of concern. Facebook, for instance, has developed a tool that uses keywords to identify and proactively report posts containing suicidal ideation. Facebook used "posts that were reported by users and the actions taken by [...] Community Operations team" to train their classifier [81]. That is, the resulting labels are based on actions (*i.e.*, human moderation decisions) that reflect determinations of content. For example, posts not about suicide are labeled as such, and no action is required. However, posts indicating suicidal ideation or intent are labeled as such, which then triggers intervention (*e.g.*, provision of resources [81]).

As this example implies, determining if content contains suicidal ideation depends on complex annotations involving myriad factors (*e.g.*, certain keywords) deemed to contribute to suicide risk [36]. These factors—such as disclosure of previous attempts, financial difficulties, and addiction [36]—may indicate suicidal ideation and describe conditions/contexts that do *not* indicate suicidal ideation. Additionally, past data may not be comprehensive, and types of content that algorithm developers did not consider may not be flagged. For example, posts may use social steganographic (*i.e.*, coded language that needs inferences) [125] techniques, such as "unalive" instead of "die", that would be missed by data collection efforts, model designers, or downstream evaluation of the tool.

Furthermore, these labels allocate attention and resources from the platform. There is potential for labels to inaccurately flag certain content (*e.g.*, posts disclosing mental health struggles, including suicidal ideation, in the context of mental health advocacy or recovery) as containing suicidal ideation, and to not flag content about suicide that failed to be classified. In these cases, these tools can exacerbate existing disparities. The labeling process—and its downstream consequences of (mis)allocated attention and interventions—intimates the potential for such algorithmic assemblages to shape conversations about suicide and perceptions of who is worthy (or not) of intervention and care. Hence, individuals who are mislabeled based on past data will need to accept the corresponding consequences. When individuals are falsely labeled as "at-risk", the visibility of their social media content may be demoted. Conversely, when individuals are incorrectly labeled as not "at-risk," they cannot receive the social support and resources that they may need. We discuss this further in Section 3.2.3. [4]

3.2.2 Stereotyping: Contextual Cues and Valencing Suicidal Expression. Considering algorithms as assemblages enables consideration of social actors as influential to automated processes. The social aspect of assemblages is perhaps especially salient to stereotyping. Assigning negative characteristics to labels, namely, applying valence to labels, constitutes stereotyping. Two types of valence may be assigned: normative and deviant. Content and accounts labeled in a normatively-valenced way receive little intervention, while content/accounts with deviantly-valenced labels receive interventions. Which labels are associated with normative and deviant is determined by social actors behind platforms (*e.g.*, moderators, developers, other users), and as such, these determinations may reinscribe social stereotypes that persist offline and reinforce social hierarchies along racial, gender, and health (*e.g.*, mental illness stigmatization) dimensions, to name a few [60]. Stereotyping may thus contribute to representational harm, aligned with [16]. As with labeling, stereotyping may be opaque to users: evidence of these elements of stigmatization often becomes legible through impacts on one's visibility on social media.

Stereotyping may also be introduced via the use of external cues as context. For example, Facebook draws on contextual cues to refine the accuracy of its ML suicide response tools, such as differentiating between sincere and sarcastic expressions (*e.g.*, "If I hear that song one more time, I'm going to kill myself") [81]. Considering the time that a post was published, the type of content, and reactions to content, Facebook's ML tools can also inform the labeling process [81]. Stereotypes arise when patterns recognized from these contextual factors unevenly target specific populations of users. For instance, adolescents may be more likely than older individuals to use hyperbolic or sarcastic expressions and more likely to post late at night/early in the morning. Flagging content from this population potentially hinders the ability of adolescent users—who may be more susceptible to particular mental health concerns than other age groups [2]—to discuss mental health concerns and their experiences of mental health, as well as seek support from peers without gaining potentially unwanted attention from the platform. Platforms like Instagram have noted that they aim to differentiate between healthy discussions and expressions of intent to self-harm [1, 3], though if and how this differentiation occurs and whether certain groups are disparately affected by ML-prompted suicide prevention interventions warrant further investigation.

3.2.3 Separation: Mediating Visibility of Content, Creators, and Tags. Once content has been labeled, content and associated accounts may undergo separation. We view separation as partly enacted through practices such as shadowbanning, flagging, content deletion, and account suspension, which may be automated and/or human-assisted processes. Separation does not only occur through demoting content/accounts labeled as deviant but also through the promotion (or not-demotion) of content labeled as normative,

[4]This analysis does not address the issue of user consent in mental health-related interventions prompted by social media platforms; whether users (can) provide meaningful consent to platforms' provision of resources or inferences about their mental health status and any associated interventions is a critical question that is beyond the scope of this paper, and one that scholars have begun to explore [157].

therefore shaping (in)visibility. As Massanari [127] notes, sorting algorithms may elevate content that reflects normative standards, such as "(white) geek masculinity" on Reddit (p. 338). These processes work in tandem to maintain a status quo and deprioritize content/accounts from or addressing marginalized communities, perspectives, and issues [103]. Flagging and deletion are two moderation possibilities once content is labeled as violating platform rules [83]. Both of these possibilities are indicative of separation. Flagging effectively separates content that is labeled as potentially deviant and marks it as requiring further moderation, while deletion prevents content from being published altogether.

Social media users often feel separation via effects on content or account visibility. Harms exacerbated by separation and visibility manipulation may at the very least be representational or allocative, depending on the type of account or content affected (*e.g.*, if the account creator depends on social media visibility for financial support). Generally, platforms may not remove content outright but alter its visibility by manipulating visibility-shaping prompts such as notifications [83]. Similarly, in the context of suicide, platforms may hide content containing certain tags that are associated with "undesirable" labels (*e.g.*, as of this writing, an Instagram search for "#suicideawareness" returns this message: "We've hidden posts for #suicideawareness to protect our community from content that may encourage behavior that can cause harm and even lead to death"). The separation of social media users as mediated in algorithmic assemblages is rather opaque to users, as practices such as shadowbanning, which affects the reach of content, are often difficult to prove [62, 67, 103, 159].

In the case of Facebook, separation occurs when an action (*e.g.*, suppressing content or prompting users with mental health resources) is activated. While Facebook may not remove posts indicating suicidal intention, separation still occurs in the sense that flagged [83] posts are routed through different moderation checks than unflagged posts. Further, there may be a social separation that occurs following content flagging. As with Instagram's blocked search terms, users searching terms related to suicide or self-harm will be (implicitly and explicitly) discouraged from connecting with others who have experienced similar struggles, and users will be unable to connect with those who have posted using terms related to suicide or self-harm. Further, research suggests that individual users may not want to be "seen" by platforms [157], and the sense that one is being monitored by a platform may discourage users from posting certain content in certain spaces, and/or use alternative terminology to avoid future detection [35].

3.2.4 Status Loss and Discrimination: Visibility as Access and Status. We view visibility and the allocation of attention (from other social media users, the platform itself) as a proxy for status on social media— it is this visibility and attention that is intervened upon by platforms [32, 57]. Users may also lose status if they lose visibility on social media. When visibility is reduced based on stereotyped labels that reinscribe hierarchies of social oppression, stereotyping and separation that limits socially disadvantaged users' access to wealth and visibility resources can lead to a structural form of discrimination via a loss of opportunity that is intertwined with status and visibility on social media. Status loss, as attached to visibility, is perhaps illustrated most clearly through cases of users who leverage social media to provide or advertise services, such as mental health advocates, influencers, and professionals. In these cases, status loss may contribute to allocative harm through financial impact [16]. Similarly, people may share or seek content on social media including Instagram and Facebook as a way to seek social support to manage their mental health [10, 40, 167, 184]. Reduced visibility of their content can mean reduced engagement and social support received from others, which we view as a type of allocative harm, also potentially leading to long-lasting effects on one's mental health, sense of community, and support-seeking behavior.

When users perceive themselves as separated—visible to or flagged by a platform due to sensitive content—they may also alter their behaviors to reduce visibility [35]. For instance, users may be "prevented from making choices that advance their preferences," such as disclosing suicidal ideation as a means of seeking support from similar others or sharing stories. Reduced visibility of such content may contribute to autonomy harm [45] – a type of representational harm [16, 105]. This perceived lack of control stands to be amplified by factors such as the perceived opacity of algorithmic tools as well as the perceived potential biases of human agents who design, operate, and participate in (*e.g.*, human-facilitated content moderation) such systems. In combination, these factors may affect users' perceptions of safety and control over conversations about suicide and mental health, as well as to whom such conversations are (in)visible, which can, in turn, contribute to the stigmatization of mental illness more broadly.

4 DISCUSSION

We examined how stigma elements [121] manifest in algorithmic assemblages in two arguably different cases (*i.e.*, "risk prediction" algorithms in higher education, as well as suicide expression and ideation detection on social media). Our analysis shows how the four stigma elements (*i.e.*, labeling, stereotyping, separation, status loss/discrimination) manifest in algorithmic assemblages, affording to attend to both social and technical elements. Through these analyses, we identify *algorithmic stigmatization* as the mechanism producing a distinct algorithmic harm (*i.e.*, *algorithmic stigma*) in addition to other harms (*i.e.*, other representational or allocative harms) along the way.

This work has implications for algorithmic systems including and beyond what we interrogated. First, we have shown how "risk-prediction" algorithms in higher education and suicide ideation prediction and detection on social media implicate algorithmic stigmatization and can impose algorithmic stigma and thus algorithmic harms. This is, in and of itself, significant – raising questions about these systems' ethical permissibility and social implications. Rather than assuming that these systems' benefits to stigmatized groups outweigh harms to them, we suggest that HCI and other scholars and practitioners designing, building, and evaluating these systems grapple and address how these systems may produce algorithmic stigma(tization).

Second, that we observe algorithmic stigmatization in both of these cases, despite their differences, demonstrates algorithmic stigmatization's relevance in a range of algorithmic assemblages beyond these cases and its utility as a theoretical concept in examining

algorithms' social and ethical implications. Researchers, practitioners, and other actors may replicate our approach to explicate where and how algorithmic stigma may manifest in their systems to begin to make informed decisions about what to do about them (*e.g.*, not build said systems, work to interrupt stigmatization). At the very least, following our approach can help elucidate where and how these systems may inflict algorithmic stigma.

Future work could examine algorithmic stigma(tization) from the perspective of individuals implicated by it (*e.g.*, students subjected to risk prediction algorithms) and explore ways to operationalize and measure said harm to further theorize algorithmic stigma(tization) across contexts.

Overall, in conceptualizing algorithmic stigmatization, we show how stigma is no longer just a social process as reviewed in Section 2.1, but also a *sociotechnical* one. In what follows, we first elaborate on the significance of algorithmic stigmatization and conclude with a discussion on where we may begin to disrupt it.

4.1 Algorithmic Stigma, Algorithmic Stigmatization, and Power

First, in applying Link & Phelan's [121] four elements of stigmatization to two cases, we rendered visible how algorithmic assemblages can produce stigma, a mechanism we call *algorithmic stigmatization*. This concept builds on the understanding that "algorithmic" refers to what "is produced by or related to an information system committed (both functionally and ideologically)" [79] – which we interpret as one entailing both functional/technical and other (*i.e.*, human/social/institutional) elements. Stigmatization – as a sociotechnical mechanism as we show – elucidates how algorithms interact with people, social processes, and institutional contexts, where algorithms are not independent entities that have power over something (or people) and thus bring about harms [33]. Rather, algorithms themselves shape and are shaped by the contexts where they are embedded, as dynamic "hybrid assemblages" and relational entities that co-constitute both humans and nonhumans [33].

Algorithmic stigmatization is an explanatory mechanism (see Figure 1) that 1) demonstrates how previously known algorithmic harms (*i.e.*, allocative and representational) can be produced and are interconnected with 2) *algorithmic stigma* (*i.e.*, stigmatization's ultimate outcome). In turn, algorithmic stigma is stigma that is mediated, perpetuated, or even created by/in algorithmic assemblages. Algorithmic stigmatization reveals how allocative and representational harms contribute to and are entangled within algorithmic stigma, but that algorithmic stigma is distinct from these harms: it occurs when *all* four stigma elements converge. While representational and allocational harms are helpful abstractions to describe a range of harms, we argue that they are inadequate to fully capture the harms caused by stigma and stigmatization as manifested in algorithmic assemblages.

Sociologists have described stigma as "a classificatory form of power" [182], "violence from above" [185], and "a bureaucratised form of violence" [50] among others. At their cores, what these conceptions of stigma as a social process have in common is their attention to where, how, by whom, and for what purpose stigma is produced. This is akin to Paton's concept of "gazing up" [147] (*i.e.*, foregrounding the stigmatizers' role, rather than the stigmatized). By applying stigma theory to algorithmic assemblages and considering algorithmic assemblages as stigmatizers in this work, we can examine algorithmic stigma as a process of power [147, 179, 180], and as "a form of governance which legitimizes the reproduction and entrenchment of inequalities and injustices" [179]. Prior work has established that sorting and structuring of social opportunity is exacerbated by algorithmic systems. Gangadharan [76], for instance, describes how digital inclusion leads to increased social sorting and increased cumulative disadvantage (following Gandy [75]), while Eubanks notes how the "digital poorhouse" acts to systematically separate, label, and limit resources [68]. Algorithmic stigma(tization), as a process of power as we have developed here, affords a theoretical lens through which to view these forms of social sorting. We suggest that algorithmic stigmatization as a theoretical concept reveals an important view into *how* stigma governs—not just socially and spatially [147, 182] as past stigma scholarship argues, but also sociotechnically.

Second, we argue that algorithmic stigmatization provides algorithmic assemblages with stigma power. Stigma power describes "instances in which stigma processes achieve the aims of stigmatizers with respect to the exploitation, management, control or exclusion of others" [120]. In the case of risk prediction algorithms in higher education, university personnel can draw on algorithmic outputs to manage and control "at-risk" students, such as through remediation or forced meetings with faculty or academic advisors. Additionally, university personnel can exclude "at-risk" students from reaping the benefits of higher education, which can take the form of remediation, encouraging students to switch to a "less demanding" major, or using risk predictions to inform budgetary matters and "minimise the potential waste of public funds spent on students who subsequently fail" [42]. In the case of social media platforms' engagement with suicidal expressions on their platforms, this may take the form of profiting from social media engagement and content [197], while providing a legal shield against being held responsible for suicide mediated or facilitated by their platforms and engagement on them. In both of these cases, algorithmic assemblages succeed in maintaining the status quo, be that in the treatment of students or social media users' mental health-related expressions and outcomes thereof.

In sum, algorithmic stigma and algorithmic stigmatization provide a theoretically-grounded tool set for both examining algorithmic systems' ethical implications (*e.g.*, harms) and for explicating mechanisms leading to algorithmic harms – of particular relevance to scholarship concerned with algorithms' societal and ethical implications across fields (*e.g.*, HCI, CSCW, FAccT). Identifying and naming algorithmic harms is a necessary first step to any possible attempt in addressing them [14].

4.2 Disrupting Algorithmic Stigmatization

Stigma is, to say the least, challenging to ameliorate. As Link and Phelan argue, it is a "persistent predicament" because "when powerful groups forcefully label and extensively stereotype a less powerful group, the range of mechanisms for achieving discriminatory outcomes is both flexible and extensive" [121] (p. 379). Algorithmic stigmatization, we argue, is going to be similarly difficult to

address – as we have shown, the processes through which algorithmic stigmatization emerges invoke a complex arrangement of social and technical elements, making it difficult to pinpoint ways to interrupt stigmatization that attend to *both* the social and technical.

Nonetheless, addressing stigmatization is essential to fostering a more just and equitable society, as is addressing algorithmic stigmatization, partly due to the outcomes stigma generates. Indeed, stigma can facilitate many outcomes, from status loss (*i.e.*, "downward placement in the status hierarchy"), to outcomes having little to do with the motivations behind stigmatization, to outcomes related to how people experiencing stigma may cope with it [121], to name a few. Additionally, algorithmic stigmatization can have both immediate and downstream effects; some effects may not occur or be observable for years. Examining algorithmic stigmatization's outcomes is a promising area for future research.

Link and Phelan [121] argue that challenging stigma should be 1) multi-faceted (to address the many mechanisms that can lead to stigma) and multi-level (to address both individual and structural discrimination), and 2) "must either change the deeply held attitudes and beliefs of powerful groups that lead to labeling, stereotyping, setting apart, devaluing, and discriminating, or it must change circumstances so as to limit the power of such groups to make their cognitions the dominant ones." (p. 381) They argue that strategies targeting one mechanism at a time are doomed to fail as "their effectiveness will be undermined by contextual factors that are left untouched by such a narrowly conceived intervention" [121]. The amelioration of algorithmic stigmatization is a massive undertaking requiring an ongoing research agenda. We hope the present work will enable this challenging yet high-impact research agenda by explicating algorithmic stigmatization. The amelioration of algorithmic stigmatization is beyond this paper's intentions and possibilities; nonetheless, in what follows we speculate on how algorithmic stigmatization may be disrupted, informed by understandings of stigma disruption strategies [48, 54, 55, 89, 119, 163] and reparative [60] approaches to addressing algorithmic harms. We suggest that future work should explore these, and other, possibilities for disrupting algorithmic stigmatization.

As reviewed above, a series of mechanisms have been proposed to mitigate algorithmic harms, largely following technical and reductionist approaches such as improving accuracy and equal representation [51, 60, 70, 87]. These approaches are often inadequate as they fail to capture the sociotechnical enablers and sources of harms [97], which are necessary to effectively do any meaningful intervention to address or mitigate harms.

While addressing algorithmic harms is notoriously difficult, scholars have proposed other kinds of approaches such as algorithmic reparation [60] which we interpret to include refusal (*i.e.*, refusing to engage with, create, or otherwise support and legitimize algorithmic systems as a form of reparation) [43, 78], participatory algorithm design [118, 143, 173], value-sensitive algorithm design [196], and examining the perspectives of social groups most adversely impacted by algorithmic systems [7, 25, 56, 157]. This last approach's reparative power rests in affording the determination of whether/when/in what contexts the use of algorithms is ethically permissible to begin with, and if there ought to be an algorithmic system, what its (un)ethical design, deployment, and use may entail. Indeed, unlike reductionist approaches, algorithmic reparation does not seek to "de-bias" algorithms. Instead, it aims for structural redress in the form of "algorithmic reform" requiring technical and social expertise [60]. For example, archivist curation, an algorithmic reparative approach, can account for the complexity in the data related to lending, hiring, and criminal sentencing, where the complex relationships between data and a variety of variables "are intractable for data practitioners alone" [60]. Curation professionals' expertise can help with the consideration of "consent, power, inclusivity, transparency, and ethics & privacy" in the collection and management of sociocultural data [102]. On the other hand, as a participatory approach to algorithm design [118, 143, 173], marginalized communities most adversely impacted by algorithmic systems of credit scoring can participate in the reform of the relationship between algorithms and social processes by co-creating algorithms and informing the values embedded in them by developers [60]. However, we note that these suggestions must serve only as a starting point in further work, rather than complete proposed solutions to solving algorithmic stigmatization.

Partial solutions for each of the four algorithmic stigmatization elements may be possible. Link and Phelan note [121] that when it comes to labeling, "the critical sociological issue is to determine how culturally created categories arise and how they are sustained...What are the social, economic, and cultural forces that maintain the focus on a particular human difference?" (p. 368) Addressing stigmatization in the creation of labels by developers of algorithms, within a multi-level, multi-faceted, and power-shifting approach, may be a first step to pinpointing where and how labeling might be disrupted. For example, a reparative approach to labeling could entail revisiting what sets of labels are or are not available to use, how they are developed, what and whose values they represent, and implications thereof. It could also entail being intentional about choices in including and excluding labels – while lack of representation can be harmful, representation can also be harmful [98]. One may also consider interrogating forgetting practices and "data silences" [136] in labeling. In light of considering forgetting practices in labeling, we suggest that reparation may include recognizing that annotators may disagree on how to label data [134]; these disagreements tend to be erased by opting for the most popular label among annotators, but as Muller and Strohmayer suggest, can reflect various subjectivities along race, gender, nationality, class, and other identity facets [136]. Regarding the stereotyping element, perhaps instead of associating a label that has always been associated with a negative value, a different value would help disrupt stigmatization.

Similar approaches towards discrimination and status loss may be feasible. Mechanisms relying on human intervention (*e.g.*, review by a human moderator) may perpetuate social biases and stigma and uphold algorithmic stigmatization. Recent tools aimed at demystifying social media bans (*e.g.*, Online Identity Help Center [5]), particularly those experienced by marginalized social media users, may aid in interrupting or challenging separation through informational assistance and identifying mechanisms for responding to stigmatizing social media moderation. This approach, while being a partial solution, could potentially shift power dynamics.

[5] https://www.oihc.org/

Alternatively, one may also draw inspiration from resistance and reparative approaches involving advocacy, collective action, and coalition building which have the potential to shift power dynamics. While traditionally, advocacy to disrupt stigma has been a task performed by marginalized impacted groups [54, 55, 163], we advocate that a reparative approach would entail advocacy by others such as civil rights groups, researchers, and community-centered organizations and in solidarity with impacted groups. Examples include civil rights advocates' efforts in a growing number of U.S. cities to ban or otherwise regulate facial recognition and other surveillance technologies that disproportionately harm marginalized groups [44, 106] or other ordinances on AI and surveillance technologies.

For example, in our second case analysis, harms furthered through the stereotyping aspect of stigmatization might be interrupted through traditional strategies like advocacy that highlights issues associated with the labels and values that are fundamental to algorithmic classification. This is complicated, however, by the opacity of algorithmic systems to affected users and lack of meaningful opportunities for contestability [138]. While platforms such as Facebook have provided insights into how their ML-driven suicide detection tools operate, greater transparency around which datasets are used to train ML tools, how labels are derived and assigned, and other factors are still warranted in interrogating and mitigating harms. These interventions' opacity complicates the ability and extent to which users can contest power structures upheld by algorithmic assemblages we described here. However, there are instances in which users have leveraged their collective experience to pressure platforms to admit to or change practices. For example, social media users pressured TikTok to admit that it suppressed content by disabled, queer, and/or fat creators [28] and called on TikTok to address the alleged shadowbanning of videos with the hashtags #BlackLivesMatter and #GeorgeFloyd [130]. Faced with similar allegations, Instagram vowed to review their practices around content suppression as it pertains to Black creators [172]. Overall, advocacy here may take the form of pushing for regulation of treating social media users' data and resulting mental health-related inferences as health data, demanding transparency of processes and outcomes, or demanding implementing meaningful informed consent and contestation processes.

A final approach with the potential to shift power is facilitating intergroup contact, another stigma disruption strategy in in-person settings (*e.g.*, [48]). Intergroup contact, in this context, may mean algorithm developers and other decision-makers in relative positions of power (compared to the stigmatized) would engage with and know about the downstream impacts of their technologies and decisions. For example, knowing what students labeled "at risk" or individuals in psychological distress think about the algorithms we examined here has the potential to contribute to how these decision-makers and developers approach problem formulation, technology design, and deployment. Presumably, such knowledge may shift values and attitudes in technology creation and deployment and as a result, shift power.

5 CONCLUSION

In this paper, we investigate what sociological conceptions of stigmatization (*i.e.*, the convergence of stigma elements: labeling, stereotyping, separating, discrimination/status loss) may explain about harms manifested in algorithmic assemblages. We apply the aforementioned four stigma elements to two algorithmic assemblages: a) "risk prediction" algorithms in higher education, and b) suicidal expression and ideation detection on social media. Through these case analyses, we contribute the novel theoretical conceptualization of *algorithmic stigmatization* as a sociotechnical mechanism that leads to an algorithmic harm we refer to as *algorithmic stigma*. This conceptualization of algorithmic stigmatization reveals how other algorithmic harms (*e.g.*, allocative and representational) contribute to and are intertwined within algorithmic stigma, but that algorithmic stigma is distinct in that it occurs when *all* four stigma elements converge. Although representational and allocational harms are insightful abstractions that can explain some harms, they fall short of thoroughly capturing stigma(tization)-related harms in algorithmic assemblages. We define algorithmic stigma as the type of stigma and algorithmic harm that is mediated, perpetuated, or sometimes created by/in algorithmic assemblages. We discuss reparative approaches' promises for beginning to disrupt algorithmic stigmatization, noting that just as stigma as a social mechanism is challenging to mitigate, so is algorithmic stigmatization as a *sociotechnical* mechanism. Nonetheless, recognizing algorithmic harms and articulating the mechanisms leading to them is a first and necessary step to begin to address algorithmic harms (including algorithmic stigma) with the promise to foster more equitable and just sociotechnical futures.

ACKNOWLEDGMENTS

We appreciate the anonymous ACs and reviewers for their encouraging and constructive feedback on this work.

REFERENCES

[1] 2019. New resources to support our community's well-being. https://newsroom.tiktok.com/en-us/new-resources-to-support-well-being
[2] 2022. Mental Illness. https://www.nimh.nih.gov/health/statistics/mental-illness
[3] Adam Mosseri. 2020. Addressing Self-Harm Content on EU Instagram | Instagram Blog. https://about.instagram.com/blog/announcements/an-important-step-towards-better-protecting-our-community-in-europe
[4] Clifford Adelman. 1998. The kiss of death? An alternative view of college remediation. *National Crosstalk* 6, 3 (1998), 11.
[5] Muhammad Adnan, Asad Habib, Jawad Ashraf, Shafaq Mussadiq, Arsalan Ali Raza, Muhammad Abid, Maryam Bashir, and Sana Ullah Khan. 2021. Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models. *IEEE Access* 9 (2021), 7519–7539. https://doi.org/10.1109/ACCESS.2021.3049446 Conference Name: IEEE Access.
[6] Nazanin Andalibi. 2020. Disclosure, Privacy, and Stigma on Social Media: Examining Non-Disclosure of Distressing Experiences. *ACM Transactions on Computer-Human Interaction* 27, 3 (May 2020), 18:1–18:43. https://doi.org/10.1145/3386600
[7] Nazanin Andalibi and Justin Buss. 2020. The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–16. https://doi.org/10.1145/3313831.3376680
[8] Nazanin Andalibi and Patricia Garcia. 2021. Sensemaking and Coping After Pregnancy Loss: The Seeking and Disruption of Emotional Validation Online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–32. https://doi.org/10.1145/3449201
[9] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. 2018. Social Support, Reciprocity, and Anonymity in Responses to Sexual Abuse Disclosures on Social Media. *ACM Transactions on Computer-Human Interaction* 25, 5 (Oct. 2018), 1–35. https://doi.org/10.1145/3234942

[10] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of #Depression. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 1485–1500. https://doi.org/10.1145/2998181.2998243

[11] Matthias C. Angermeyer and Herbert Matschinger. 2005. Labeling—stereotype—discrimination: An investigation of the stigma process. *Social Psychiatry and Psychiatric Epidemiology* 40, 5 (May 2005), 391–395. https://doi.org/10.1007/s00127-005-0903-4

[12] Matthias C. Angermeyer, Herbert Matschinger, Bruce G. Link, and Georg Schomerus. 2014. Public attitudes regarding individual and structural discrimination: Two sides of the same coin? *Social Science & Medicine* 103 (Feb. 2014), 60–66. https://doi.org/10.1016/j.socscimed.2013.11.014

[13] Kimberly E. Arnold and Matthew D. Pistilli. 2012. Course signals at Purdue: using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. ACM, Vancouver British Columbia Canada, 267–270. https://doi.org/10.1145/2330601.2330666

[14] Jennifer Bard. 2020. Developing a Legal Framework for Regulating Emotion AI. https://doi.org/10.2139/ssrn.3680909

[15] Pınar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2021. Person, Human, Neither: The Dehumanization Potential of Automated Image Tagging. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Virtual Event USA, 357–367. https://doi.org/10.1145/3461702.3462567

[16] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem with Bias: From Allocative to Representational Harms in Machine Learning.

[17] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671. https://heinonline.org/HOL/Page?handle=hein.journals/calr104&id=695&div=&collection=

[18] Howard S. Becker. 1963. *Outsiders: Studies in the sociology of deviance*. Free Press Glencoe, Oxford, England. Pages: x, 179.

[19] Jeri Mullins Beggs, John H. Bantham, and Steven Taylor. 2008. Distinguishing the factors influencing college students' choice of major. *College Student Journal* 42, 2 (June 2008), 381–395. https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=01463934&v=2.1&it=r&id=GALE%7CA179348418&sid=googleScholar&linkaccess=abs Publisher: Project Innovation Austin LLC.

[20] Ruha Benjamin. 2019. Assessing risk, automating racism. *Science* 366, 6464 (Oct. 2019), 421–422. https://doi.org/10.1126/science.aaz3873

[21] Ruha Benjamin. 2020. Race After Technology: Abolitionist Tools for the New Jim Code. *Social Forces* 98, 4 (June 2020), 1–3. https://doi.org/10.1093/sf/soz162

[22] Cynthia L. Bennett and Os Keyes. 2020. What is the point of fairness? disability, AI and the complexity of justice. *ACM SIGACCESS Accessibility and Computing* 125 (March 2020), 5:1. https://doi.org/10.1145/3386296.3386301

[23] Johannes Berens, Kerstin Schneider, Simon Görtz, Simon Oster, and Julian Burghoff. 2019. Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. 11, 3 (2019), 41.

[24] Jacqueline Bichsel and EDUCAUSE Center for Applied Research. 2012. *Analytics in higher education: benefits, barriers, progress, and recommendations*. EDUCAUSE Center for Applied Research, Louisville, Colo. http://net.educause.edu/ir/library/pdf/ERS1207/ers1207.pdf OCLC: 807289562.

[25] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (Feb. 2021), 100205. https://doi.org/10.1016/j.patter.2021.100205

[26] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158 ISSN: 2642-9381.

[27] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[28] Elena Botella. 2019. TikTok admits it suppressed videos by disabled, queer, and fat creators. https://slate.com/technology/2019/12/tiktok-disabled-users-videos-suppressed.html

[29] Martha Cecilia Bottia, Roslyn Arlin Mickelson, Cayce Jamil, Kyleigh Moniz, and Leanne Barry. 2021. Factors Associated With College STEM Participation of Racially Minoritized Students: A Synthesis of Research. *Review of Educational Research* 91, 4 (Aug. 2021), 614–648. https://doi.org/10.3102/00346543211012751 Publisher: American Educational Research Association.

[30] Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. MIT Press. Google-Books-ID: xHlP8WqzizYC.

[31] Keffrelyn D. Brown. 2006. *Mapping risks in education: Conceptions, contexts and complexities*. Ph.D. The University of Wisconsin - Madison, United States – Wisconsin. https://www.proquest.com/docview/304977319/abstract/F5898C2ED3A3430FPQ/1 ISBN: 9780542887093.

[32] Taina Bucher. 2012. Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. *New Media & Society* 14, 7 (Nov. 2012), 1164–1180. https://doi.org/10.1177/1461444812440159 Publisher: SAGE Publications.

[33] Taina Bucher. 2018. *If...Then: Algorithmic Power and Politics*. Oxford University Press. Google-Books-ID: u_pdDwAAQBAJ.

[34] Scott Burris. 2006. Stigma and the law. *The Lancet* 367, 9509 (Feb. 2006), 529–531. https://doi.org/10.1016/S0140-6736(06)68185-3 Publisher: Elsevier.

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

[36] Stevie Chancellor, Steven A. Sumner, Corinne David-Ferdon, Tahirah Ahmad, and Munmun De Choudhury. 2021. Suicide Risk and Protective Factors in Online Support Forum Posts: Annotation Scheme Development and Validation Study. *JMIR Mental Health* 8, 11 (Nov. 2021), e24471. https://doi.org/10.2196/24471 Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada.

[37] Pamara F. Chang and Rachel V. Tucker. 2022. Assistive Communication Technologies and Stigma: How Perceived Visibility of Cochlear Implants Affects Self-Stigma and Social Interaction Anxiety. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (April 2022), 77:1–77:16. https://doi.org/10.1145/3512924

[38] Yan Chen, Qinghua Zheng, Shuguang Ji, Feng Tian, Haiping Zhu, and Min Liu. 2020. Identifying at-risk students based on the phased prediction model. *Knowledge and Information Systems* 62, 3 (March 2020), 987–1003. https://doi.org/10.1007/s10115-019-01374-x

[39] Shaan Chopra, Rachael Zehrung, Tamil Arasu Shanmugam, and Eun Kyoung Choe. 2021. Living with Uncertainty and Stigma: Self-Experimentation and Support-Seeking around Polycystic Ovary Syndrome. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445706

[40] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. In *Eighth International AAAI Conference on Weblogs and Social Media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8075

[41] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (April 2020), 82–89. https://doi.org/10.1145/3376898

[42] Kwok Tai Chui, Dennis Chun Lok Fung, Miltiadis D. Lytras, and Tin Miu Lam. 2020. Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior* 107 (June 2020), 105584. https://doi.org/10.1016/j.chb.2018.06.032

[43] Marika Cifor, Patricia Garcia, TL Cowan, Jasmine Rault, Tonia Sutherland, Anita Say Chan, Jennifer Rode, Anna Lauren Hoffmann, Niloufar Salehi, and Lisa Nakamura. 2019. Feminist Data Manifest-No. https://www.manifestno.com/home

[44] Giovanni Circo. 2020. PROJECT GREENLIGHT DETROIT: EVALUATION REPORT. (2020), 94.

[45] Danielle Keats Citron and Daniel J. Solove. 2021. Privacy Harms. https://doi.org/10.2139/ssrn.3782222

[46] Matthew Clair. 2018. Stigma. *Core Concepts in Sociology* (2018).

[47] Patricia Hill Collins and Sirma Bilge. 2020. *Intersectionality*. John Wiley & Sons. Google-Books-ID: fyrfDwAAQBAJ.

[48] Jonathan E. Cook, Valerie Purdie-Vaughns, Ilan H. Meyer, and Justin T. A. Busch. 2014. Intervening within and across levels: A multilevel approach to stigma and public health. *Social Science & Medicine* 103 (Feb. 2014), 101–109. https://doi.org/10.1016/j.socscimed.2013.09.023

[49] A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 864–876. https://doi.org/10.1145/3531146.3533150

[50] Vickie Cooper and David Whyte (Eds.). 2017. *The Violence of Austerity*. Pluto Press. https://doi.org/10.2307/j.ctt1pv8988

[51] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. http://arxiv.org/abs/1808.00023 arXiv:1808.00023 [cs].

[52] Patrick Corrigan. 2004. How stigma interferes with mental health care. *American Psychologist* 59, 7 (Oct. 2004), 614–625. https://doi.org/10.1037/0003-066X.59.7.614

[53] Patrick W. Corrigan, Benjamin G. Druss, and Deborah A. Perlick. 2014. The Impact of Mental Illness Stigma on Seeking and Participating in Mental Health Care. *Psychological Science in the Public Interest* 15, 2 (Oct. 2014), 37–70. https://doi.org/10.1177/1529100614531398 Publisher: SAGE Publications Inc.

[54] Patrick W. Corrigan and Mandy W. M. Fong. 2014. Competing perspectives on erasing the stigma of illness: What says the dodo bird? *Social Science & Medicine* 103 (Feb. 2014), 110–117. https://doi.org/10.1016/j.socscimed.2013.05.027

[55] Patrick W. Corrigan, Jonathon E. Larson, Patrick J. Michaels, Blythe A. Buchholz, Rachel Del Rossi, Malia Javier Fontecchio, David Castro, Michael Gause, Richard Krzyżanowski, and Nicolas Rüsch. 2015. Diminishing the self-stigma of mental illness by coming out proud. *Psychiatry Research* 229, 1 (Sept. 2015), 148–154. https://doi.org/10.1016/j.psychres.2015.07.053

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

[57] Kelley Cotter. 2019. Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. *New Media & Society* 21, 4 (April 2019), 895–913. https://doi.org/10.1177/1461444818815684 Publisher: SAGE Publications.

[58] Kimberle Crenshaw. 1990. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review* 43 (1990), 1241. https://heinonline.org/HOL/Page?handle=hein.journals/stflr43&id=1257&div=&collection=

[59] David Danks and Alex John London. 2017. Algorithmic Bias in Autonomous Systems. (2017), 7.

[60] Jenny L. Davis, Apryl Williams, and Michael W. Yang. 2021. Algorithmic reparation. *Big Data & Society* 8, 2 (July 2021), 20539517211044808. https://doi.org/10.1177/20539517211044808 Publisher: SAGE Publications Ltd.

[61] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society* 8, 2 (July 2021), 20539517211035955. https://doi.org/10.1177/20539517211035955 Publisher: SAGE Publications Ltd.

[62] Brooke Erin Duffy. 2020. Algorithmic precarity in cultural work. *Communication and the Public* 5, 3-4 (Sept. 2020), 103–107. https://doi.org/10.1177/2057047320959855 Publisher: SAGE Publications.

[63] Dustin T. Duncan and Ichiro Kawachi. 2018. *Neighborhoods and Health.* Oxford University Press. Google-Books-ID: SZNODwAAQBAJ.

[64] Upol Ehsan, Ranjit Singh, Jacob Metcalf, and Mark Riedl. 2022. The Algorithmic Imprint. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul Republic of Korea, 1305–1317. https://doi.org/10.1145/3531146.3533186

[65] Karin van Es, Daniel Everts, and Iris Muis. 2021. Gendered language and employment Web sites: How search algorithms can cause allocative harm. *First Monday* (July 2021). https://doi.org/10.5210/fm.v26i8.11717

[66] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users' Behavior Around Them in Rating Platforms. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (May 2017), 62–71. https://doi.org/10.1609/icwsm.v11i1.14898 Number: 1.

[67] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300724

[68] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Publishing Group. Google-Books-ID: pn4pDwAAQBAJ.

[69] Sina Fazelpour and David Danks. 2021. Algorithmic bias: Senses, sources, solutions. *Philosophy Compass* 16, 8 (2021), e12760. https://doi.org/10.1111/phc3.12760 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/phc3.12760.

[70] Sina Fazelpour and Zachary C. Lipton. 2020. Algorithmic Fairness from a Non-ideal Perspective. http://arxiv.org/abs/2001.09773 arXiv:2001.09773 [cs, stat].

[71] Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination.* MIT Press. Google-Books-ID: 8ZiWDwAAQBAJ.

[72] Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. Value Sensitive Design and Information Systems. In *Early engagement and new technologies: Opening up the laboratory*, Neelke Doorn, Daan Schuurbiers, Ibo van de Poel, and Michael E. Gorman (Eds.). Springer Netherlands, Dordrecht, 55–95. https://doi.org/10.1007/978-94-007-7844-3_4

[73] Daniel James Fuchs. 2018. The Dangers of Human-Like Bias in Machine-Learning Algorithms. 2 (2018), 15.

[74] Liza Gak, Seyi Olojo, and Niloufar Salehi. 2022. The Distressing Ads That Persist: Uncovering The Harms of Targeted Weight-Loss Ads Among Users with Histories of Disordered Eating. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–23. https://doi.org/10.1145/3555102 arXiv:2204.03200 [cs].

[75] Oscar H. Gandy. 2016. *: Engaging Rational Discrimination and Cumulative Disadvantage.* Routledge, London. https://doi.org/10.4324/9781315572758

[76] Seeta Peña Gangadharan. 2012. Digital inclusion and data profiling. *First Monday* (April 2012). https://doi.org/10.5210/fm.v17i5.3821

[77] Alex Gano. 2017. Disparate Impact and Mortgage Lending: A Beginner's Guide. *University of Colorado Law Review* 88 (2017), 1109. https://heinonline.org/HOL/Page?handle=hein.journals/ucollr88&id=1145&div=&collection=

[78] Patricia Garcia, Tonia Sutherland, Marika Cifor, Anita Say Chan, Lauren Klein, Catherine D'Ignazio, and Niloufar Salehi. 2020. No: Critical Refusal as Feminist Data Practice. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing.* ACM, Virtual Event USA, 199–202. https://doi.org/10.1145/3406865.3419014

[79] Tarleton Gillespie. 2016. Algorithm. In *2. Algorithm.* Princeton University Press, 18–30. https://doi.org/10.1515/9781400880553-004

[80] Erving Goffman. 1986. *Stigma: Notes on the management of spoiled identity* (1st touchstone ed. ed.). Simon & Schuster, New York.

[81] Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. 2018. Ethics and Artificial Intelligence: Suicide Prevention on Facebook. *Philosophy & Technology* 31, 4 (Dec. 2018), 669–684. https://doi.org/10.1007/s13347-018-0336-0

[82] Miguel Angel González-Torres, Rodrigo Oraa, Maialen Arístegui, Aranzazu Fernández-Rivas, and Jose Guimon. 2007. Stigma and discrimination towards people with schizophrenia and their family members. *Social Psychiatry and Psychiatric Epidemiology* 42, 1 (Jan. 2007), 14–23. https://doi.org/10.1007/s00127-006-0126-3

[83] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (Jan. 2020), 2053951719897945. https://doi.org/10.1177/2053951719897945 Publisher: SAGE Publications Ltd.

[84] Louis F. Graham, Mark B. Padilla, William D. Lopez, Alexandra M. Stern, Jerry Peterson, and Danya E. Keene. 2016. Spatial Stigma and Health in Postindustrial Detroit. *International Quarterly of Community Health Education* 36, 2 (Jan. 2016), 105–113. https://doi.org/10.1177/0272684X15627800 Publisher: SAGE Publications Inc.

[85] Sylvia Sims Gray. 2013. Framing "at risk" students: Struggles at the boundaries of access to higher education. *Children and Youth Services Review* 35, 8 (Aug. 2013), 1245–1251. https://doi.org/10.1016/j.childyouth.2013.04.011

[86] Ben Green. 2021. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. (2021), 33.

[87] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* ACM, Barcelona Spain, 19–31. https://doi.org/10.1145/3351095.3372840

[88] Amanda L. Griffith. 2008. Determinants of Grades, Persistence and Major Choice for Low-Income and Minority Students. (May 2008). https://ecommons.cornell.edu/handle/1813/74602 Accepted: 2020-11-17T16:57:44Z.

[89] Rita Guerra, Margarida Rebelo, Maria B. Monteiro, and Samuel L. Gaertner. 2013. Translating Recategorization Strategies Into an Antibias Educational Intervention. *Journal of Applied Social Psychology* 43, 1 (2013), 14–23. https://doi.org/10.1111/j.1559-1816.2012.00976.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1559-1816.2012.00976.x.

[90] Ian Hacking. 2004. Between Michel Foucault and Erving Goffman: between discourse in the abstract and face-to-face interaction. *Economy and Society* 33, 3 (Aug. 2004), 277–302. https://doi.org/10.1080/0308514042000225671 Publisher: Routledge _eprint: https://doi.org/10.1080/0308514042000225671.

[91] Oliver L. Haimson, Justin Buss, Zu Weinger, Denny L. Starks, Dykee Gorrell, and Briar Sweetbriar Baron. 2020. Trans Time: Safety, Privacy, and Content Warnings on a Transgender-Specific Social Media Site. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–27. https://doi.org/10.1145/3415195

[92] Emma Halliday, Jennie Popay, Rachel Anderson de Cuevas, and Paula Wheeler. 2020. The elephant in the room? Why spatial stigma does not receive the public health attention it deserves. *Journal of Public Health* 42, 1 (Feb. 2020), 38–43. https://doi.org/10.1093/pubmed/fdy214

[93] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* ACM, Barcelona Spain, 501–512. https://doi.org/10.1145/3351095.3372826

[94] Stacey Hannem and Chris Bruckert. 2012. *Stigma Revisited: Implications of the Mark.* University of Ottawa Press. Google-Books-ID: rGykDAAAQBAJ.

[95] Mark L. Hatzenbuehler. 2016. Structural Stigma and Health Inequalities: Research Evidence and Implications for Psychological Science. *The American psychologist* 71, 8 (Nov. 2016), 742–751. https://doi.org/10.1037/amp0000068

[96] Todd F. Heatherton. 2003. *The Social Psychology of Stigma.* Guilford Press.

[97] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7 (June 2019), 900–915. https://doi.org/10.1080/1369118X.2019.1573912 Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2019.1573912.

[98] Anna Lauren Hoffmann. 2021. Terms of inclusion: Data, discourse, violence. *New Media & Society* 23, 12 (Dec. 2021), 3539–3556. https://doi.org/10.1177/

1461444820958725 Publisher: SAGE Publications.

[99] Lily Hu and Issa Kohler-Hausmann. 2020. What's sex got to do with machine learning?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 513. https://doi.org/10.1145/3351095.3375674

[100] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 560–575. https://doi.org/10.1145/3442188.3445918

[101] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901

[102] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 306–316. https://doi.org/10.1145/3351095.3372829

[103] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. 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 (Oct. 2021), 305:1–305:44. https://doi.org/10.1145/3476046

[104] Atoosa Kasirzadeh and Andrew Smart. 2021. The Use and Misuse of Counterfactuals in Ethical Machine Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 228–236. https://doi.org/10.1145/3442188.3445886

[105] Kate Crawford. 2017. The trouble with bias.

[106] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and P. M. Krafft. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 45–55. https://doi.org/10.1145/3351095.3372874

[107] Jared Katzman, Solon Barocas, Su Lin Blodgett, Kristen Laird, Morgan Klaus Scheuerman, and Hanna Wallach. 2023. Representational Harms in Image Tagging. *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence* (2023), 5.

[108] David Kay and Professor Charles Oppenheim. [n. d.]. Vol.1 No6.: Legal, Risk and Ethical Aspects of Analytics in Higher Education. ([n. d.]), 30.

[109] Gunay Kazimzade and Milagros Miceli. 2020. Biased Priorities, Biased Outcomes: Three Recommendations for Ethics-oriented Data Annotation Practices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 71. https://doi.org/10.1145/3375627.3375809

[110] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 88:1–88:22. https://doi.org/10.1145/3274357

[111] Os Keyes, Zoë Hitzig, and Mwenza Blell. 2021. Truth from the machine: artificial intelligence and the materialization of identity. *Interdisciplinary Science Reviews* 46, 1-2 (April 2021), 158–175. https://doi.org/10.1080/03080188.2020.1840224

[112] Andrew E. Krumm, R. Joseph Waddington, Stephanie D. Teasley, and Steven Lonn. 2014. A Learning Management System-Based Early Warning System for Academic Advising in Undergraduate Engineering. In *Learning Analytics: From Research to Practice*, Johann Ari Larusson and Brandon White (Eds.). Springer, New York, NY, 103–119. https://doi.org/10.1007/978-1-4614-3305-7_6

[113] Anna Kruse and Rob Pongsajapan. 2012. Student-Centered Learning.

[114] Himabindu Lakkaraju, Everaldo Aguiar, Carl Shan, David Miller, Nasir Bhanpuri, Rayid Ghani, and Kecia L. Addison. 2015. A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Sydney NSW Australia, 1909–1918. https://doi.org/10.1145/2783258.2788620

[115] Anders Larrabee Sønderlund, Emily Hughes, and Joanne Smith. 2019. The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology* 50, 5 (2019), 2594–2618. https://doi.org/10.1111/bjet.12720 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.12720.

[116] Celeste Lawson, Colin Beer, Dolene Rossi, Teresa Moore, and Julie Fleming. 2016. Identification of 'at risk' students using learning analytics: the ethical dilemmas of intervention strategies in a higher education institution. *Educational Technology Research and Development* 64, 5 (Oct. 2016), 957–968. https://doi.org/10.1007/s11423-016-9459-0

[117] Susan Leavy, Barry O'Sullivan, and Eugenia Siapera. 2020. Data, Power and Bias in Artificial Intelligence. http://arxiv.org/abs/2008.07341 arXiv:2008.07341 [cs].

[118] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–35. https://doi.org/10.1145/3359283

[119] Mark Levine, Amy Prosser, David Evans, and Stephen Reicher. 2005. Identity and Emergency Intervention: How Social Group Membership and Inclusiveness of Group Boundaries Shape Helping Behavior. *Personality and Social Psychology Bulletin* 31, 4 (April 2005), 443–453. https://doi.org/10.1177/0146167204271651 Publisher: SAGE Publications Inc.

[120] Bruce G. Link and Jo Phelan. 2014. Stigma power. *Social Science & Medicine* 103 (Feb. 2014), 24–32. https://doi.org/10.1016/j.socscimed.2013.07.035

[121] Bruce G. Link and Jo C. Phelan. 2001. Conceptualizing Stigma. *Annual Review of Sociology* 27, 1 (Aug. 2001), 363–385. https://doi.org/10.1146/annurev.soc.27.1.363

[122] Bruce G. Link, Jo C. Phelan, and Mark L. Hatzenbuehler. 2014. Stigma and Social Inequality. In *Handbook of the Social Psychology of Inequality*, Jane D. McLeod, Edward J. Lawler, and Michael Schwalbe (Eds.). Springer Netherlands, Dordrecht, 49–64. https://doi.org/10.1007/978-94-017-9002-4_3

[123] Owen H. T. Lu, Jeff C. H. Huang, Anna Y. Q. Huang, and Stephen J. H. Yang. 2017. Applying learning analytics for improving students engagement and learning outcomes in an MOOCs enabled collaborative programming course. *Interactive Learning Environments* 25, 2 (Feb. 2017), 220–234. https://doi.org/10.1080/10494820.2016.1278391

[124] Juan F. Maestre. 2020. Conducting HCI Research on Stigma. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing (CSCW '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 129–134. https://doi.org/10.1145/3406865.3418364

[125] Alice E Marwick and danah boyd. 2014. Networked privacy: How teenagers negotiate context in social media. *New Media & Society* 16, 7 (Nov. 2014), 1051–1067. https://doi.org/10.1177/1461444814543995 Publisher: SAGE Publications.

[126] Ozma Masood. 2009. *At risk: the racialized student marked for educational failure*. Ph. D. Dissertation. Library and Archives Canada = Biblioth??que et Archives Canada, Ottawa. ISBN: 9780494399248 OCLC: 649886910.

[127] Adrienne Massanari. 2017. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19, 3 (March 2017), 329–346. https://doi.org/10.1177/1461444815608807 Publisher: SAGE Publications.

[128] Martha Maxwell. 1980. *Improving student learning skills /*. Jossey-Bass,. https://eduq.info/xmlui/handle/11515/9515 Accepted: 2015-11-06T12:21:48Z Artwork Medium: Ressource physique Interview Medium: Ressource physique.

[129] Martha Maxwell. 1997. The Dismal State of Required Developmental Reading Programs: Roots, Causes and Solutions. (June 1997). https://eric.ed.gov/?id=ED415501

[130] Megan McCluskey. 2020. Black TikTok Creators Say Their Content Is Being Suppressed | Time. https://time.com/5863350/tiktok-black-creators/

[131] Samuel Messick. 1996. Validity and washback in language testing. *Language Testing* 13, 3 (Nov. 1996), 241–256. https://doi.org/10.1177/026553229601300302 Publisher: SAGE Publications Ltd.

[132] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 735–746. https://doi.org/10.1145/3442188.3445935

[133] Ilan H. Meyer. 2003. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: Conceptual issues and research evidence. *Psychological Bulletin* 129, 5 (2003), 674–697. https://doi.org/10.1037/0033-2909.129.5.674 Place: US Publisher: American Psychological Association.

[134] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 115:1–115:25. https://doi.org/10.1145/3415186

[135] Nanlir Sallau Mullah and Wan Mohd Nazmee Wan Zainon. 2021. Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access* 9 (2021), 88364–88376. https://doi.org/10.1109/ACCESS.2021.3089515 Conference Name: IEEE Access.

[136] Michael Muller and Angelika Strohmayer. 2022. Forgetting Practices in the Data Sciences. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3517644

[137] Zane Muller. 2019. Algorithmic Harms to Workers in the Platform Economy: The Case of Uber. *Columbia Journal of Law and Social Problems* 53 (2019), 167. https://heinonline.org/HOL/Page?handle=hein.journals/collsp53&id=179&div=&collection=

[138] Deirdre K. Mulligan, Daniel Kluttz, and Nitin Kohli. 2019. Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions. https://doi.org/10.2139/ssrn.3311894

[139] Rob Nixon. 2011. *Slow Violence and the Environmentalism of the Poor*. Harvard University Press. Google-Books-ID: e3jDDwAAQBAJ.

[140] Mutale Nkonde. 2019. Automated Anti-Blackness: Facial Recognition in Brooklyn, New York. (2019), 7.

[141] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press. https://doi.org/10.18574/nyu/9781479833641.001.0001 Publication Title: Algorithms of Oppression.

[142] Devah Pager. 2003. The Mark of a Criminal Record. *Amer. J. Sociology* 108, 5 (March 2003), 937–975. https://doi.org/10.1086/374403 Publisher: The University of Chicago Press.

[143] Joon Sung Park, Karrie Karahalios, Niloufar Salehi, and Motahhare Eslami. 2022. Power Dynamics and Value Conflicts in Designing and Maintaining Socio-Technical Algorithmic Processes. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (March 2022), 1–21. https://doi.org/10.1145/3512957

[144] Richard Parker and Peter Aggleton. 2003. HIV and AIDS-related stigma and discrimination: a conceptual framework and implications for action. *Social Science* (2003), 12.

[145] Frank Pasquale. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information.* Harvard University Press. Google-Books-ID: TumaBQAAQBAJ.

[146] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* ACM, Atlanta GA USA, 39–48. https://doi.org/10.1145/3287560.3287567

[147] Kirsteen Paton. 2018. Beyond legacy: Backstage stigmatisation and 'trickle-up' politics of urban regeneration. *The Sociological Review* 66, 4 (July 2018), 919–934. https://doi.org/10.1177/0038026118777449 Publisher: SAGE Publications Ltd.

[148] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (Nov. 2021), 100336. https://doi.org/10.1016/j.patter.2021.100336

[149] Shenira A. Perez. 2019. *Quantifying the Effects of the 'At-risk' Label: Exploring the Deficit-oriented Labeling Experiences of Low-income, First-generation College Students of Color.* Ph.D. Boston College, United States – Massachusetts. https://www.proquest.com/docview/2289587090/abstract/96398F2C745C44A1PQ/1 ISBN: 9781085695701.

[150] Sidney Perkowitz. 2021. The Bias in the Machine: Facial Recognition Technology and Racial Disparities. *MIT Case Studies in Social and Ethical Responsibilities of Computing* Winter 2021 (Feb. 2021). https://doi.org/10.21428/2c646de5.62272586 Publisher: MIT Schwarzman College of Computing.

[151] Margaret Placier. 1996. The Cycle of Student Labels in Education: The Cases of Culturally Deprivedl Disadvantaged and at Risk. *Educational Administration Quarterly* 32, 2 (April 1996), 236–270. https://doi.org/10.1177/0013161X96032002004 Publisher: SAGE Publications Inc.

[152] Devin G. Pope and Justin R. Sydnor. 2011. Implementing Anti-discrimination Policies in Statistical Profiling Models. *American Economic Journal: Economic Policy* 3, 3 (Aug. 2011), 206–231. https://doi.org/10.1257/pol.3.3.206

[153] T. T. A. Putri, S. Sriadhi, R. D. Sari, R. Rahmadani, and H. D. Hutahaean. 2020. A comparison of classification algorithms for hate speech detection. *IOP Conference Series: Materials Science and Engineering* 830, 3 (April 2020), 032006. https://doi.org/10.1088/1757-899X/830/3/032006 Publisher: IOP Publishing.

[154] Emily Rauscher and William Elliott III. 2014. The Effect of Wealth Inequality on Higher Education Outcomes: A Critical Review. *Sociology Mind* 04, 04 (2014), 282–297. https://doi.org/10.4236/sm.2014.44029

[155] Lauren Rhue. 2018. Racial Influence on Automated Perceptions of Emotions. https://doi.org/10.2139/ssrn.3281765

[156] Rashida Richardson, Jason Schultz, and Kate Crawford. 2019. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. https://papers.ssrn.com/abstract=3333423

[157] Kat Roemmich and Nazanin Andalibi. 2021. Data Subjects' Conceptualizations of and Attitudes Toward Automatic Emotion Recognition-Enabled Wellbeing Interventions on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 308:1–308:34. https://doi.org/10.1145/3476049

[158] Edward G. Rozycki. 2004. At-Risk Students: What Exactly Is the Threat? How Imminent Is It? *Educational Horizons* 82, 3 (2004), 174–179. https://www.jstor.org/stable/42926497 Publisher: [Sage Publications, Ltd., Phi Delta Kappa International].

[159] Laura Savolainen. 2022. The shadow banning controversy: perceived governance and algorithmic folklore. *Media, Culture & Society* 44, 6 (Sept. 2022), 1091–1109. https://doi.org/10.1177/01634437221077174 Publisher: SAGE Publications Ltd.

[160] Devansh Saxena and Shion Guha. 2022. Algorithms in the Daily Lives of Child-Welfare Caseworkers: Harms to Practice, Agency, and Street-Level Decision-Making. https://doi.org/10.2139/ssrn.4077245

[161] Graham Scambler. 2009. Health-related stigma. *Sociology of Health & Illness* 31, 3 (2009), 441–455. https://doi.org/10.1111/j.1467-9566.2009.01161.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9566.2009.01161.x.

[162] Morgan Klaus Scheuerman, Madeleine Pape, and Alex Hanna. 2021. Auto-essentialization: Gender in automated facial analysis as extended colonial project. *Big Data & Society* 8, 2 (July 2021), 20539517211053712. https://doi.org/10.1177/20539517211053712 Publisher: SAGE Publications Ltd.

[163] Toni Schmader, Alyssa Croft, Jessica Whitehead, and Jeff Stone. 2013. A Peek Inside the Targets' Toolbox: How Stigmatized Targets Deflect Discrimination by Invoking a Common Identity. *Basic and Applied Social Psychology* 35, 1 (Jan. 2013), 141–149. https://doi.org/10.1080/01973533.2012.746615 Publisher: Routledge _eprint: https://doi.org/10.1080/01973533.2012.746615.

[164] Vanessa Scholes. 2016. The ethics of using learning analytics to categorize students on risk. *Educational Technology Research and Development* 64, 5 (Oct. 2016), 939–955. https://doi.org/10.1007/s11423-016-9458-1

[165] Nete Schwennesen. 2019. Algorithmic assemblages of care: imaginaries, epistemologies and repair work. *Sociology of Health & Illness* 41, S1 (2019), 176–192. https://doi.org/10.1111/1467-9566.12900 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9566.12900.

[166] Nick Seaver. 2019. Knowing Algorithms. In *DigitalSTS: A Field Guide for Science & Technology Studies.* 568.

[167] Ellen Selkie, Victoria Adkins, Ellie Masters, Anita Bajpai, and Daniel Shumer. 2020. Transgender Adolescents' Uses of Social Media for Social Support. *Journal of Adolescent Health* 66, 3 (March 2020), 275–280. https://doi.org/10.1016/j.jadohealth.2019.08.011

[168] George Siemens and Phil Long. 2011. Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review* 46, 5 (2011), 30. Publisher: EDUCAUSE.

[169] Ellen Simpson and Bryan Semaan. 2021. For You, or For"You"? Everyday LGBTQ+ Encounters with TikTok. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (Jan. 2021), 252:1–252:34. https://doi.org/10.1145/3432951

[170] Sharon Slade and Paul Prinsloo. 2013. Learning Analytics: Ethical Issues and Dilemmas. *American Behavioral Scientist* 57, 10 (Oct. 2013), 1510–1529. https://doi.org/10.1177/0002764213479366 Publisher: SAGE Publications Inc.

[171] Tom Slater. 2017. Territorial Stigmatization: Symbolic Defamation and the Contemporary Metropolis. In *The SAGE Handbook of New Urban Studies.* SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road London EC1Y 1SP, 111–125. https://doi.org/10.4135/9781412912655.n8

[172] Todd Spangler. 2020. Instagram to Review Whether Its Practices and Policies 'Suppress Black Voices'. https://variety.com/2020/digital/news/instagram-suppress-black-voices-review-algorithmic-bias-1234635850/

[173] Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Ken Holstein, Zhiwei Steven Wu, and Haiyi Zhu. 2022. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders. In *2022 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Seoul Republic of Korea, 1162–1177. https://doi.org/10.1145/3531146.3533177

[174] Terrell L. Strayhorn. 2015. Factors Influencing Black Males' Preparation for College and Success in STEM Majors: A Mixed Methods Study. *Western Journal of Black Studies* 39, 1 (2015), 45–63. https://www.proquest.com/docview/1688657629/abstract/2DD905147AEB4A58PQ/1 Num Pages: 19 Place: Pullman, United States Publisher: Washington State University Press.

[175] Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization.* ACM, – NY USA, 1–9. https://doi.org/10.1145/3465416.3483305

[176] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (May 2013), 44–54. https://doi.org/10.1145/2447976.2447990

[177] Rahel Süß. 2022. The right to disidentification: Sovereignty in digital democracies. *Constellations* n/a, n/a (2022). https://doi.org/10.1111/1467-8675.12626 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-8675.12626.

[178] Zeynep Tufekci. 2015. Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency. *Colorado Technology Law Journal* 13 (2015), 203. https://heinonline.org/HOL/Page?handle=hein.journals/jtelhtel13&id=227&div=&collection=

[179] Imogen Tyler. 2013. *Revolting Subjects: Social Abjection and Resistance in Neoliberal Britain.* Bloomsbury Publishing. Google-Books-ID: DQI1EAAAQBAJ.

[180] Imogen Tyler. 2018. Resituating Erving Goffman: From Stigma Power to Black Power. *The Sociological Review* 66, 4 (July 2018), 744–765. https://doi.org/10.1177/0038026118777450 Publisher: SAGE Publications Ltd.

[181] Imogen Tyler. 2020. *Stigma: The Machinery of Inequality.* Bloomsbury Publishing. Google-Books-ID: Y_80EAAAQBAJ.

[182] Imogen Tyler and Tom Slater. 2018. Rethinking the sociology of stigma. *The Sociological Review* 66, 4 (July 2018), 721–743. https://doi.org/10.1177/0038026118777425 Publisher: SAGE Publications Ltd.

[183] David L. Vogel, Rachel L. Bitman, Joseph H. Hammer, and Nathaniel G. Wade. 2013. BRIEF REPORT Is Stigma Internalized? The Longitudinal Impact of Public Stigma on Self-Stigma.

[184] Piper Vornholt and Munmun De Choudhury. 2021. Understanding the Role of Social Media–Based Mental Health Support Among College Students: Survey and Semistructured Interviews. *JMIR Mental Health* 8, 7 (July 2021), e24512. https://doi.org/10.2196/24512 Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada.

[185] Loïc Wacquant. 2008. *Urban Outcasts: A Comparative Sociology of Advanced Marginality.* Polity.

[186] Loïc Wacquant, Tom Slater, and Virgílio Borges Pereira. 2014. Territorial Stigmatization in Action. *Environment and Planning A: Economy and Space* 46, 6 (June 2014), 1270–1280. https://doi.org/10.1068/a4606ge Publisher: SAGE Publications Ltd.

[187] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring Representational Harms in Image Captioning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 324–335. https://doi.org/10.1145/3531146.3533099

[188] Xin Wang, Peng Zhao, and Xi Chen. 2020. Fake news and misinformation detection on headlines of COVID-19 using deep learning algorithms. *International Journal of Data Science* 5, 4 (Jan. 2020), 316–332. https://doi.org/10.1504/IJDS.2020.115873 Publisher: Inderscience Publishers.

[189] T. Wangsness and J. Franklin. 1966. "Algorithm" and "formula". *Commun. ACM* 9, 4 (April 1966), 243. https://doi.org/10.1145/365278.365286

[190] Amy C. Watson, Patrick Corrigan, Jonathon E. Larson, and Molly Sells. 2007. Self-Stigma in People With Mental Illness. *Schizophrenia Bulletin* 33, 6 (Nov. 2007), 1312–1318. https://doi.org/10.1093/schbul/sbl076

[191] Lindsay Weinberg. 2022. Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *Journal of Artificial Intelligence Research* 74 (May 2022), 75–109. https://doi.org/10.1613/jair.1.13196

[192] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in Social Media: Definition, Manipulation, and Detection. *ACM SIGKDD Explorations Newsletter* 21, 2 (Nov. 2019), 80–90. https://doi.org/10.1145/3373464.3373475

[193] Daphna Yeshua-Katz and Ylva Hård af Segerstad. 2020. Catch 22: The Paradox of Social Media Affordances and Stigmatized Online Support Groups. *Social Media + Society* 6, 4 (Oct. 2020), 2056305120984476. https://doi.org/10.1177/2056305120984476 Publisher: SAGE Publications Ltd.

[194] Felice Yeskel. 2008. Coming to Class: Looking at Education through the Lens of Class Introduction to the Class and Education Special Issue. *Equity & Excellence in Education* 41, 1 (Feb. 2008), 1–11. https://doi.org/10.1080/10665680701793428

[195] Elana Zeide. 2021. The Silicon Ceiling: How Algorithmic Assessments Construct an Invisible Barrier to Opportunity. https://www.elanazeide.com/post/the-silicon-ceiling-how-algorithmic-assessments-construct-an-invisible-barrier-to-opportunity

[196] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 194:1–194:23. https://doi.org/10.1145/3274463

[197] Shoshana Zuboff. 2018. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (1st ed.).