Computational approaches to understanding sensitive mental health content are necessary as social media platforms grow in both their importance and size. Although most use online communities for pro-social and pro-health behaviors, other communities encourage *high-risk health behaviors,* such as suicidal ideation, pro-eating disorder, and opioid addiction – behaviors that can be life-threatening. These behaviors can have paradoxical "contagion" effects - spreading negative emotions and behaviors to those without a disorder [1], yet also providing therapeutic benefits from participation [2]. Platforms and communities grapple with how to manage these behaviors because of their entanglements with physical well-being, platform guidelines and community policies, governance, and moral imperatives for interventions.

**High-risk mental health behaviors and the communities that support them are a rich domain to explore current tensions in machine learning applied to sociotechnical problems.** New research deploys automated tools built with machine learning and computational linguistics to identify and intervene in unwanted and dangerous behaviors in online communities. However, data-driven approaches without considering stakeholders and context oversimplify the complexities of mental disorders and unique interactions of these communities with individual decision-making and platform governance. As AI and machine learning emerge as fields with popular research and media attention, the field of computing must reconcile the development of powerful algorithms for prediction and inference alongside the applications of these techniques to problems of great societal importance.

**My research builds and critically examines human-centered algorithms for high-risk health behaviors as a lens to empirically study tensions in human-centered machine learning.** I blend methodological contributions from machine learning, natural language processing, and data science with collaborations and theoretical insights from fields like clinical psychology and communication. My research builds rigorous applications of data-driven methods on millions of posts on social media that accurately identify and assess dangerous mental health behaviors. Additionally, I innovate in how to conduct *human-centered* machine learning, an approach that deliberately refocuses technological design and implementation on the needs of humans, communities, and stakeholders. Drawing on my background in Communication and Media Studies, I draw from Agre's "critical technical" approach [3], interrogating the practices and ethics in my field to both reflexively inform my work and advocate for thoughtful approaches in machine learning.

**In this work, I bring data-driven methods to Human Computer Interaction and Social Computing and a human-centered approach to Data Science and Machine Learning.** This is exemplified by a research agenda that has produced 13 publications, 11 first author, in premier venues such as CHI, CSCW, and FAT*. My first-authored work has been awarded four Best Paper Honorable Mention awards [4, 5, 6, 7], and produced sustained collaborations with clinical and industry stakeholders (Centers for Disease Control and Prevention; Columbia University; University of Rochester; Yahoo!; Microsoft Research).

Drawing on these interests, **my research agenda will define and advocate for a paradigm of human-centered machine learning that provides rigorous and thoughtful approaches for tough questions about algorithms and human behavior.** My future work will continue this trajectory of incorporating stakeholders and human-centered insights into understanding, developing, and deploying algorithmic solutions for problems in online communities. Below, I discuss my work as it relates to technical innovation in applied machine learning and empirical and pragmatic critique of practices within the field. I then describe my plans for research, as I work towards stronger standards in human-centered machine learning, health, and online communities.

## Technical Components of Human-Centered Machine Learning

In my prior work, I have deeply explored the pro-eating disorder (pro-ED) community as an extended case study of high-risk health behaviors. This includes data-driven methods to study online behavior as well as integrating human insights into algorithm design and implementation. My work allows for us to understand and predict mental illness behaviors in social media and understand the latent impacts of these communities on platforms.

***Technical Innovations in Applied Machine Learning:*** Pro-ED content is multimodal and combines image, text, and hashtag information to signal its intent - consider the subtleties of a thin fashion model tagged to #fashion compared to #thighgap. Current automated approaches lack this contextual sensitivity to identifying subtle behaviors that convey pro-ED intent. I addressed this gap by deploying deep learning to identify multimodal pro-ED content that violate Tumblr's platform guidelines [8]. I combined convolutional neural networks for image analysis through AlexNet and word embeddings that pass to a deep neural network for prediction with performance at 89% accuracy. I tuned our input data, annotation schemes, and error analysis to the needs of moderators on the platform and demonstrated multiple methods for integrating this system into current moderation practices.

Additional research examines how technical innovation can make sense of behaviors in online communities. For mental disorders, individuals' mental illness severity (MIS) varies over time, and this can influence their own predispositions to participate in communities as well as provide key insights for developing strategic interventions. I designed a novel statistical method combining Latent Dirichlet Allocation (LDA) probability distributions with novice/clinician annotations to infer MIS in over 26 million Instagram posts [4]. I then used these markers of severity to forecast MIS up to seven months in advance. I found, alarmingly, that MIS is on the rise on these platforms (+13%/year) and made recommendations for assessing community well-being. This paper was awarded a Best Paper Honorable Mention at CSCW [4].

***Human-Centered Approaches Directed to Algorithm Design:*** To contextualize data-driven findings and techniques, I supplement machine learning approaches with human-centered design of algorithms. This involves methods innovations through the study design process to better understand our area of interest.

One persistent challenge in social media research is obtaining *ground truth,* or accurate labels to give to prediction systems. Domain experts give reliable and accurate labels, but their efforts are difficult to scale across millions of posts. By sustaining collaborations with these stakeholders, my work develops novel methods of labeling ground truth to thoughtfully scale human labeling tasks [4, 5, 8, 9]. In the same project about MIS, I worked with clinical psychologists to annotate 150 topics with a score of their MIS and apply these ratings to label 26 million posts [4]. This effectively scaled up the ratings of the domain experts to larger sets of data than what can normally be managed by humans and provided a method for tracking MIS changes across the community. I have also used expert annotations to rate data for machine learning classes [5,8], tuning the outputs of models with expert advice and judgment [4,11,12], and conducting error analyses on mistakes made by the classifier [6,8,11]. These approaches are tempered with human sensitivity to interpret the outputs of machine learning, placing them in clinical [8] and community contexts [10, 12].

Another challenge in understanding these high-risk behaviors is identifying communities who avoid detection. For pro-ED, one method of evading platform-enforced content moderation is by changing hashtag spellings to avoid hard bans on words - #thighgap is banned on Instagram, so the community moved to the semantically similar #thyghgapp. I designed an innovative combination of algorithmic snowball sampling of hashtags with human curation to track these lexical changes and comprehensively study the pro-ED community [5]. Starting with known pro-ED tags on Instagram, I iteratively sampled tens of millions of posts to identify potential pro-ED hashtags [5]. These hashtags were then curated by domain experts and organized into semantically similar hashtag roots and variants (#thighgap is considered the "root" of #thyghgapp). Through my sustained collaborations with domain experts, I have worked to improve the study design and operationalization process for human-centered machine learning.

## Research in the Ecosystem of Human-Centered Machine Learning

Human-centered approaches to machine learning encompass more than just technical innovation alone. These approaches include a deliberate refocusing on the needs and goals of individuals, communities, and groups to drive machine learning innovation and application. By prioritizing the needs of humans and communities, we can gain deep and nuanced insights into human behavior that inform design and interventions, and fundamentally better represent the phenomenon of interest.

***Reflecting on Practices and Ethics:*** Recently, I have critically interrogated the methods, practices, and ethics in my work to both reflexively inform my practices as well as advocate for these approaches in AI and machine learning. This new line of research works to empirically identify gaps and build frameworks to encourages data scientists and quantitative researchers to more critically consider the risks of predictive analysis, participants' needs, and how to handle these in responsible, rigorous ways.

In one project appearing in PACM-HCI/CSCW, I explore the ways that human-centered machine learning frames the human research subject in peer-reviewed scientific papers. Using thematic discourse analysis on a dataset of 55 published papers, I found that the research community describes these individuals and communities as both the benefits of machine learning and the object and target of dehumanization. This paper was awarded a Best Paper Honorable Mention [7]. In another work under submission at *npj digital medicine*, I studied how interdisciplinary researchers establish construct validity and replicability in their work on the same topic area, discovering that these publications identify divergent and sometimes inconsistent methodological choices. Currently, I am beginning to work with social media ethicist to study the local norms and practices for research ethics in human-centered machine learning.

***Platform Governance:*** Platforms struggle with moderating high-risk health communities, in part because they evade detection and communities believe that moderation is not effective in curbing this behavior. To tackle this challenge, I devised the first quantitative study to investigate the impacts of moderation on pro-ED communities on Instagram. I extracted 713 variant hashtags of 17 "root" tags, all of which received content moderation in April 2012 by Instagram [5]. Using computational linguistics to analyze 8 million posts, I found that the pro-ED communities had adopted non-standard lexical variants to circumvent these restrictions. These new communities showed 15-30% more likes/comments on posts and discussed more dangerous and toxic topics compared to the communities on the original hashtags. I concluded that moderation efforts were not successful in reducing pro-ED content on Instagram. This paper was awarded a Best Paper Honorable Mention at CSCW [5]. Follow-up work explored the roles of self-censorship on platforms by self-deleting content [10] and developing robust and powerful classifiers to assist moderators in identifying behavior [7]. By using human-centered algorithms to understand online communities, we gain precise and nuanced insights into human behavior, and can begin to develop compassionate and effective behavioral interventions.

## Future Research in Human-Centered Machine Learning

My growing research agenda will examine the sociotechnical systems around high-risk health behaviors in online social platforms, and work to establish a paradigm of rigorous and compassionate work in human-centered machine learning. I am excited to investigate the following areas of research going forward:

***Opioid Addiction and Recovery Communities**:* Extending beyond the pro-ED community, I am interested in studying other examples of high-risk health behavior in online communities. One example are communities that advocate for clinically unverified treatments for opioid addiction, like unregulated drugs and off-label use of known substances. Expanding on recent publications in CHI [11], I am working with a clinical substance abuse expert to extract dosage information for these alternative treatments. This would be valuable for medical and clinical audiences, as there is very little understanding of these substances in the current literature. I also am looking to interviewing and engaging with these communities in participatory algorithmic design to better identify their behaviors and provide meaningful support. I have recently received a small grant for this project through Northwestern's Undergraduate Research program, and I anticipate using this to build towards applying to larger agencies like NIH/NIMH as well as foundations interested in solving the opioid crisis in the US.

***Ground Truth, Crowd Work, and Suicidal Ideation:*** As I mentioned, there are many challenges of acquiring gold standard labels for datasets in machine learning and health. I am interested in the ways that non-experts/lay people could be "trained" to annotate information for machine learning goals, thereby allowing machine learning to more easily scale to varied needs like in public health, crisis intervention, or clinical use. In collaboration with the CDC, we are designing a study to teach crowd workers how to annotate for risk factors and protective factors in suicidal ideation on Reddit crisis communities. Our long-term goal is to identify different ways to support the development of scalable, multi-use machine learning methods for mental health and the accompanying datasets needed to classification. A proof-of-concept is in its early stages, and I envision a larger study of developing better annotation systems funded by NSF or industry funding opportunities, where ground truth data is a bottleneck in many corporate infrastructures.

***Designing for Differential Interventions:*** Given my interest into human-centered approaches to computation, augmenting computational insights with more human-centered methods is essential to understanding peoples' lived experiences and creating interventions to facilitate compassionate interactions. During my postdoctoral fellowship at Northwestern, I am exploring and designing for differential, contextual health needs in online communities. This will be a multi-stage study, examining why people come to and engage with high-risk health communities through interviews, computational temporal analyses, and design exercises. This project is one example of an expanded pipeline for human-centered machine learning work, placing machine learning and computational approaches in discussions with other methods to better support the needs of online communities.

## References

[1] Jeanne B. Martin. "The development of ideal body image perceptions in the United States." *Nutrition Today* 45, no. 3 (2010): 98-110.

[2] Tobit Emmens and Andy Phippen. "Evaluating Online Safety Programs." *Harvard Berkman Center for Internet and Society.[23 July 2011]* (2011).

[3] Phil Agre. "Toward a critical technical practice: Lessons learned in trying to reform AI". In Bowker, G., Star, S., Turner, W., and Gasser, L., eds, *Social Science, Technical Systems and Cooperative Work: Beyond the Great Divide, Erlbaum* (1997).

[4] **Stevie Chancellor**, Zhiyuan (Jerry) Lin, Erica Goodman, Stephanie Zerwas, and Munmun De Choudhury. (2016). *Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities*. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing. CSCW 2016.

[5] **Stevie Chancellor**, Jessica 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 and Social Computing. CSCW 2016

[6] **Stevie Chancellor** and Scott Counts. (2018*). Measuring Employment Demand Using Internet Search Data*. In Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems. CHI 2018.

[7] **Stevie Chancellor**, Eric P.S. Baumer, and Munmun De Choudhury. (2019) *Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media.* (2019). Accepted in the PACM – Human Computer Interaction, to be presented at CSCW 2019

[8] **Stevie Chancellor,** Yannis Kalantidis, Jessica Pater, Munmun De Choudhury, and David A. Shamma (2017). *Multimodal Classification of Moderated Online Pro-Eating Disorder Content*. In Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems. CHI 2017.

[9] **Stevie Chancellor,** Tanushree Mitra, and Munmun De Choudhury. (2016). *Recovery Amid Pro-Anorexia: Analysis of Recovery in Social Media*. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI 2016.

[10] **Stevie Chancellor,** Zhiyuan (Jerry) Lin, and Munmun De Choudhury. (2016). *"This Post Will Just Get Taken Down": Characterizing Removed Pro-Eating Disorder Social Media Content.* In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI 2016.

[11] **Stevie Chancellor,** George Nitzburg, Andrea Hu, Francisco Zampieri, and Munmun De Choudhury. (2019). *Discovering Alternative Treatments for Opioid Use Recovery in Social Media*. In Proceedings of the ACM Conference on Human Factors in Computing Systems. CHI 2019.

[12] **Stevie Chancellor**, Andrea Hu, and Munmun De Choudhury. (2018). *Norms Matter: Contrasting Social Support Around Behavior Change in Online Weight Loss Communities*. In Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems. CHI 2018.

[13] **Stevie Chancellor,** Michael Birnbaum, Eric Caine, Vince Silenzio, and Munmun De Choudhury. (2019). *Ethical Tensions in Inferring Mental Health States from Social Media: Questions and Calls to Action*. In Proceedings of the Conference on ACM Fairness, Accountability, and Transparency (FAT*).