RESEARCH STATEMENT

SHAGUN JHAVER

HCI/CSCW RESEARCHER WORKING ON CONTENT MODERATION, HUMAN-MACHINE COLLABORATION, AND THE FUTURE OF WORK.

Social media platforms like Facebook, Twitter and Reddit have an increasingly essential role in regulating the flow of information online. They make day-to-day decisions about what content is allowed to stay on their site and what is removed, and how users' information is shared with their networks and the larger public. Platforms make these decisions by implementing complex socio-technical systems, often referred to as content moderation. How these systems are designed and how they allow content to flow have direct implications not only for what counts as free speech online but for democracy itself. Yet, over the past few years, media sources have frequently reported that platforms fail to remove disturbing content, block content that is important to be circulated in the public sphere, and promote content that is related to conspiracy theories. The moderation systems on these platforms are also notoriously opaque, revealing little about how they operate. Therefore, scrutinizing the internal processes of these systems is crucial to understanding their limitations and to building better solutions.

My research builds a foundation for **designing fair and efficient content moderation systems**. I have contributed in-depth descriptions of how moderation systems on Twitter and Reddit are constituted and how they affect platform owners, content moderators, and end-users. I have also evaluated the effectiveness of different moderation strategies on the health of these platforms. By attending to the viewpoints and practices of different stakeholders, my research offers a holistic view of content moderation. This allows me to identify opportunities for improvements that are both meaningful for the end-users as well as feasible for the community managers to implement.

I routinely engage in **mixed-methods** research, drawing on relevant methodologies for the questions I find most pressing to explore. In my thesis, I have drawn from a wide variety of data sources, ranging from yearlong participant observations and interviews with dozens of users to hundreds of survey responses and millions of social media posts. For analyzing my data, I have used quantitative methods such as regression analyses and topic modeling as well as qualitative methods such as thematic analysis and grounded theory. I value learning insights from each data source and using them to inform the exploration of other data sources, finally integrating my findings into a more complete picture.

My work as a PhD student has been published in prestigious HCI venues such as TOCHI, CHI, CSCW and ICWSM. It has also received two Best Paper Awards (at CSCW [1] and ICWSM [2]), one Honorable Mention Award (at CSCW [3]), and been featured in Editor's Spotlight in TOCHI [4]. Moreover, my research has received attention in mainstream venues such as The Washington Post, Al Jazeera, and New Scientist.

RESEARCH CONTRIBUTIONS

My dissertation research consists of studies around four important aspects of content moderation: (1) Moderating against online harassment, (2) Enacting content moderation, (3) Designing for fairness and transparency in content moderation, and (4) Evaluating moderation mechanisms.

Moderating Online Harassment

One of the primary goals of content moderation is to detect and remove instances of online harassment. A fundamental challenge in combating online harassment is that there is no standard definition of what online harassment entails. In order to conceptualize online harassment, I adopted a primarily qualitative approach. I studied Twitter and Reddit users who claimed to have suffered online harassment as well as users who were accused of perpetrating online harassment [4, 5]. Drawing on my interviews with users who had experienced online harassment, I curated a typology of behavioral patterns that participants described as manifestations of online harassment. This **empirically grounded typology** offers community managers an important resource to identify instances of problematic behaviors and moderate such behaviors. Further, through analyzing the boundaries between free speech and online harassment in this research, I contributed a theoretical model to understand perceptions of controversial speech.

The problem of online harassment is particularly prevalent on the social media site Twitter. I studied the strategies that users deploy to protect themselves from harassment on this site. For this work, I focused on Twitter blocklists, a third-party blocking mechanism developed by volunteer users on Twitter to address online harassment [4]. Blocklists allow users to pre-emptively block with a few clicks thousands of accounts on a computationally generated list of block-worthy accounts. I conducted semi-structured interviews with 14 users who had subscribed to blocklists and a separate group of 14 users who were blocked on these lists. My findings show that users are not adequately protected from harassment, and at the same time, many people feel they are blocked unnecessarily. Building upon these findings, I offered **design solutions to address the needs of harassment victims while ensuring that users are not blocked unfairly**. I also highlighted the need for designers to focus not just on creating *blocking solutions* but also *understanding mechanisms* that allow users with differing ideologies to interact without fear of being abused.

Enacting Content Moderation

My research agenda also focuses on understanding *how* content moderators regulate the online communities they moderate. I have volunteered as a Reddit moderator for several Reddit communities over the past four years. I also conducted semi-structured interviews with moderators of five large Reddit communities [6]. I found that moderators rely on a variety of creative automated solutions that address the unique regulation requirements of their communities. For this project, I focused on one of the most popular automated tools, called Automoderator (or Automod), that is now offered to all Reddit moderators. My findings show that Automod reduces the time-consuming work and emotional labor required of human moderators by automatically removing large volumes of inappropriate content. However, the use of Automod requires moderators to develop new technical skills, defend against deliberate avoidance of Automod filters by bad actors, and reverse mistakes made by Automod. My findings also reveal the deficiencies of Automod in making decisions that require it to be attuned to the sensitivities in cultural context or to the nuances in linguistic cues. Building on this case study, I offered **theoretical and practical guidelines for how human-machine collaboration should be enacted to attain efficient content moderation** on social media websites.

Designing for Fairness and Transparency in Content Moderation

In light of the growing public and media concerns about how platforms curate content online, I have evaluated platform decisions in the framework of *Fairness, Accountability, Transparency and Ethics (FATE)* model.

I began exploring **what *fairness in content moderation* means from the perspectives of end-users** by conducting a survey of 907 Reddit users whose posts had been recently removed [3]. In this survey, I asked users how they perceived fairness of the post removal and whether they would post again on the community. Using a mixed-methods approach for data analysis, this study contributed insights for how moderators can motivate users to become productive community members. In another study, I conducted a large-scale analysis of **how transparency in moderation affects the future behavior** of social media users [1]. I focused on a specific aspect of transparency – the messages that provide users an explanation of why their post was removed. Using a sample of 32 million Reddit posts, I characterized the removal explanations that are provided to Redditors and linked them to measures of subsequent user behaviors. My regression analyses showed that offering explanations for content removals reduces the odds of future post removals. Additionally, explanations provided by human moderators did not have a significant advantage over explanations provided by bots for reducing future post removals. Building upon my findings, I proposed design solutions that can promote the efficient use of explanation mechanisms, reflecting on how automated moderation tools can contribute to this space.

Evaluating the Effectiveness of Moderation Mechanisms

In response to the growing complaints about content moderation, social media platforms have implemented novel strategies to improve content curation and reduce offensive behavior on their sites. My recent work has begun evaluating these strategies using large-scale data analyses.

When Reddit ***quarantines*** a community, accessing that community displays a warning that requires users to explicitly opt-in to viewing its content. My collaborators and I sought to **understand how this moderation strategy affects the radicalization of social media users on hate groups**. To this end, I evaluated the effects of quarantining the *r/TheRedPill* subreddit, a community known to often post misogynistic content [7]. Using interrupted time series regression analysis and bootstrapping tests that controlled for Reddit-wide trends, I found that quarantining subreddits helps "control" online radicalization in the following ways: (1) It acts as a barrier to entry, making it harder to recruit newcomers to these extremist groups, and (2) It reduces the levels of activity amongst existing users who regularly participated within these groups before the quarantine. These findings suggest that quarantining can be a surprisingly effective compromise between censoring offensive speech and preserving freedom of expression. As a follow-up to this work, I am currently investigating the effects of another moderation strategy, called "deplatforming," which refers to a permanent ban of disruptive public figures on Twitter.

FUTURE WORK

Examining Content Curation Strategies

While the current scholarship on content moderation has explored some key aspects, we have only scratched the surface of this complex, multi-layered practice. Because of a rapidly growing research, media and public interest in this topic, I view content moderation emerging as an important sub-discipline in Social Computing research. I see myself at the forefront of this sub-discipline, and I consider myself equipped to lead research in this space.

My previous work has evaluated several moderation approaches and interventions. Continuing this line of research, I will study the effectiveness of graduated sanctions as a moderation strategy. That is, I will examine whether a mild but certain punishment is more effective in deterring misbehavior than a severe but uncertain punishment. My prior work shows that online harassment disproportionately affects underrepresented groups such as minorities and LGBT users [4]. Going forward, I will explore how different approaches to addressing online harassment affect these groups. I also plan to study the appeal procedures on social media websites that allow end-users to contest moderation decisions that they find unfair. This line of research will provide empirically grounded guidelines to **improve the efficiency and legitimacy of content moderation**.

I will also go beyond social media platforms to investigate how **content curation strategies on other digital platforms** affect end-users. My previous work on Airbnb search results is an example of how semi-transparent, algorithmic content curation can increase the emotional labor of sharing economy workers [8]. Advancing this line of research, I will investigate how algorithmic content curation affects end-users who own content on other platforms like Amazon.com, Yelp, YouTube, and dating sites. Prior research in this area has largely focused on transparency of curation algorithms. I will evaluate how other ethical principles like consistency and competency of content curation algorithms affect users' perceptions of fairness of curation decisions. I will also study how users' perceived level of control over content curation outcomes influences their attitudes. This work will contribute recommendations for designing content curation algorithms in ways that **reduce the anxiety and uncertainty among end-users and promote their sense of fairness.**

Incorporating Ethical Principles in Designing Human-Algorithm Interactions

My prior work has focused on effectively integrating human and computer intelligence in content moderation systems [6]. Going forward, I am interested in studying algorithm-in-the-loop decision-making in settings such as hiring, justice and work evaluation. While debates around algorithmic decisions have emphasized the accuracy and fairness of algorithmic models, relatively little research has examined how algorithms influence human decision-making. Building algorithm-in-the loop systems that can reliably produce ethical outcomes requires considering not just the model accuracy but the full sociotechnical contexts in which people and algorithms interact. Therefore, I will evaluate decision-making systems where humans use algorithmic aids using the core principles of FATE [1, 3] as well as other theories of fairness and justice. I will identify the biases and problematic aspects of these systems. Additionally, my work will seek to understand the processes by which these problems come about and how to remedy them. This research will contribute **ethical standards and design insights for developing effective and responsible human-machine collaborations.**

Understanding and Addressing Socioeconomic Inequalities in Modern Labor Markets

My dissertation work has shown how inefficient approaches to content moderation disproportionately impacts under-represented groups [3, 4]. However, there are a multitude of other ways in which internet-based technologies can contribute to socioeconomic inequalities. During my internship at Microsoft Research, I used Bing search queries for skill development to characterize differences in the amount and type of skills likely being developed in different parts of the US [9]. I also quantified how those differences align with measures of economic growth. My findings showed that internet-based skill development contributes to economic disparities. Expanding this line of research, I will use skills and employment data from resumes, job postings, search engines and social media websites to derive insights into labor dynamics with spatial and temporal granularity. These insights could lead to policy interventions to cope with shifting economic trends as well as help municipalities and governments address local labor needs. In a recent study on skill development of Airbnb hosts, I found that hosts rely on Facebook Groups to learn from other hosts and progress from being Airbnb novices to becoming expert hosts [10]. Building upon this research, I will implement and evaluate participatorily designed social computing systems that support the communication and learning needs of sharing economy workers, especially those living in underserved communities. This line of research will **address the socio-economic and geographic inequalities in modern labor markets** and contribute to HCI discussions of future of work.

Conclusion

As a researcher and educator, I am committed to building a strong community of scholars, volunteer moderators and designers in the field of content moderation. As part of this commitment, I am co-leading a workshop at CSCW that is aimed at welcoming new researchers and industry partners and establishing a set of ethical principles to inform research practices in this field. Moving forward, I am excited to build on my prior research on content moderation, human-algorithm interactions, and the future of work. My broader research agenda is to contribute to positive social change by informing national policies and technical designs of digital systems through a user-centered approach. I believe that there is an urgent need for digital platforms to incorporate new standards of fairness, transparency and accountability in their operations. My experiences so far have put me in a unique position to meaningfully contribute towards this goal.

REFERENCES

[1] **Shagun Jhaver**, Amy Bruckman, and Eric Gilbert. Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit. *In Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 3(150), 2019.
Best Paper Award.

[2] Munmun De Choudhury, **Shagun Jhaver**, Benjamin Sugar, and Ingmar Weber. Social Media Participation in an Activist Movement for Racial Equality. In *Tenth International AAAI Conference on Web and Social Media (ICWSM)*, 2016.
Best Paper Award.

[3] **Shagun Jhaver**, Darren Scott Appling, Amy Bruckman, and Eric Gilbert. "Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit. *In Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 3(192), 2019.
Best Paper Honorable Mention Award.

[4] **Shagun Jhaver**, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online Harassment and Content Moderation: The Case of Blocklists. *In ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):12:1–12:33, March 2018.
Featured in Editor's Spotlight.

[5] **Shagun Jhaver**, Larry Chan, and Amy Bruckman. The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action. *First Monday*, 23(2), 2018.

[6] **Shagun Jhaver**, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *In ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):31:1–31:35, July 2019.

[7] **Shagun Jhaver***, Eshwar Chandrasekharan*, Amy Bruckman, and Eric Gilbert. The Quarantine of r/TheRedPill: Efficacy of a Reddit Intervention in Limiting Misogyny and Radicalization. *Submitted to the ACM CHI Conference on Human Factors in Computing Systems (CHI).* (* co-primary authors), 2020.

[8] **Shagun Jhaver**, Yoni Karpfen, and Judd Antin. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. In *Proceedings of the 2018 ACM CHI Conference on Human Factors in Computing Systems (CHI)*, pages 421:1–421:12. ACM, 2018.

[9] **Shagun Jhaver**, Justin Cranshaw, and Scott Counts. Measuring Professional Skill Development in US Cities Using Internet Search Queries. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 13, pages 267–277, 2019.

[10] Maya Holikatti, **Shagun Jhaver**, and Neha Kumar. Learning to Airbnb by Engaging in Online Communities of Practice. *In Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 3(228), 2019.