Research Statement

Hussam Habib

Online platforms have transformed our relationship with information. Algorithmic filtering, optimizing for engagement, and content curation are some of the ways through which platforms have become integral in our lives. Platforms, such as Google, Facebook, and YouTube, have transformed how we access information, interact with society, and express ourselves, yet the role they play in societal and political dysfunction is unclear (1). As studies find symptoms such as political polarization (2), misinformation (3, 4), and filter bubbles (5), the question of how the fundamental systems of online platforms shape our societal and individual behaviors becomes more pressing.

I research how algorithmic and design aspects of platforms influence user behavior. My long-term research goals are to A) investigate how platforms influence user behavior and B) understand how platforms interpret user inputs and behavior. Understanding (A) how the design of systems within platforms shapes user behavior can bring insights into the harms of platforms. Investigating (B) the behavior of platforms—how they interpret user signals and serve relevant content—is essential for controlling them. In my work, I achieve these goals using two separate frameworks: Observational studies from the perspective of the user to understand how their behavior is influenced and experimental studies from the perspective of the platform to characterize their behavior and response to user behavior.

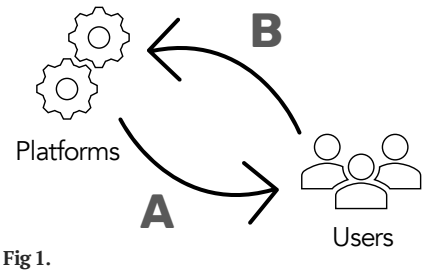

Fig 1.
A. How do platforms influence user behavior?
B. How are user inputs interpreted by platforms?

How do platforms influence us?

From the vantage point of the users, I perform observational studies on large-scale digital traces of users on online platforms. The goal of these studies is to measure, monitor, and track user behavior, which is mined using a combination of computational techniques and social science theories. This user behavior data brings insight into user beliefs and ideologies, crucially, how they changed over time on the platform. I validate relationships between user behavior and platform design, using statistical and machine learning analyses, to establish the potential harms of platforms.

How do users adopt problematic behavior? CSCW 22 • ICWSM 23
To explore this question, I investigate how users on online platforms, like Reddit, develop and exhibit radical ideologies (6). This follows from an interest in understanding how user digital traces, more specifically their language, can reveal significant information about the psychological and ideological state of the user's mind. Crucially, as shown by [cite], users exhibit specific traits in their language as they adopt problematic and radical ideologies. As platforms expose users to new ideas, this work investigates how users respond to radical ideologies—specifically radical misogyny. To this end, we track 17,000 users over 68 months on Reddit to record their interactions and behavioral changes. We find interaction with radical users, regardless of where it happens, and the influence of community feedback, positive or negative, to be key in shaping the adoption of problematic ideologies. Notably, Users parrot radical traits linked with the ideologies of the radical users they interact with. Taken together, this study emphasizes the importance of timely and complete moderation of problematic users and communities to prevent the spread of harmful ideologies—a takeaway that reoccurs throughout my other works.

In our next study, we explore how the posts—specifically their language, theme, source, and topic—correlate with problematic ideologies and behavior. To this end, we perform an exploratory study on the Facebook posts of victims of COVID-19 who proclaimed their anti-vax and COVID-19 beliefs. (7). Using computer vision techniques to interpret and characterize posts made by the victims, our analysis reveals the overwhelming politicization of the COVID-19 narrative within the posts, which often include trusted parties and figures as sources or references. Findings from both of our studies, underlining the failures of platforms and communities within, provide perspectives into how to mitigate these harms while providing insights into design and administrative decisions for fundamentally removing these pathways. Both of our works unanimously point towards the overwhelming influence of non-intervened dangerous communities, calling first for a need for improved intervention techniques.

How do communities adopt problematic behavior? ICWSM 22 • ICWSM 22
This question was of great interest as we observed the insurgence of radical far-right ideologies within communities online. I was interested in exploring how communities *evolve* and adopt problematic behavior. In this work, I constructed community embeddings that represented the content and user base of the community for a particular month. This allowed community dynamics to be captured and track how subreddits (communities on Reddit) evolved (8). Mainly, we found subreddits to be extremely dynamic and constantly changing regularly. Even more so, we found communities that were later banned because of their problematic behaviors (through violation of Reddit's content policy) evolved with significantly different patterns. Using machine learning techniques and feature engineering, that yield interpretable and understandable predictions, we investigate

this anomaly. The features served as hypotheses for the reason for the decline of communities' health. Our results show that – more than the adoption of problematic language and hate speech – the insurgence of users from a previously banned community was more likely to damage community health. To assist administrators in monitoring community health, we leveraged our predictive tool to create a proactive flagging tool. We test our flagging tool in a real-world environment, using continuous learning, where it proactively flags communities on average nine months before their actual ban. Our tool allows administrators, moderators, and the community itself to acknowledge the declining health of the community and work towards overcoming it.

Moderation has been considered a crucial step towards a healthier platform – as also outlined by my prior works. However, historically it proved to be a *make or break* for many platforms. Moderation is expensive and resource-demanding; therefore it is often at a crossroads with a platform's economic goals. To then understand ***why*** platforms moderate, we investigate the circumstances and influences around the banning of communities on Reddit (9). In our work, we score communities across Reddit on their toxicity and hate speech (a metric acting as Reddit's content policy's proxy). The time series of community's toxicity score with their bans implies a lack of timely and comprehensive intervention strategy from Reddit. This leads us to hypothesize external negative media attention towards the community is driving Reddit to intervene. We perform a mediation analysis on extracted instances of negative attention using natural language processing and find a complete mediation. This suggests the interventions are influenced by the negative attention from external sources. Although alarming and concerning, these findings corroborate the fact that platforms are rational and economic entities that optimize for profits and financial shareholders. These insights drive us to further explore the behaviors of platforms, in response to user inputs, to further investigate the relationship between users and platforms.

How do platforms interpret user inputs?

On the other end, from the vantage point of the platform to study how platforms interpret user signals and respond accordingly. I perform experimental studies using user agents to emulate user behavior. Because of the complex nature of algorithms and limited access to platforms' internals, I use a network tomography approach to uncover relationships between the user input and platform output to build an understanding of the processes within. To this end, I use web mining, software engineering, and simulation techniques to instrument browsers and deploy sock puppets on a large scale to emulate users.

How does Google amplify cognitive biases? *Under review*

Over the last two decades, the process through which individuals seek information has changed significantly. Algorithmic retrieval and ranking of information have dominated access to information as platforms like Google become ubiquitous. Yet, the artifacts and emergent effects of adopting algorithmic systems within the modern information-seeking processes are not fully understood. In this work, we seek to study the end-to-end information-seeking processes for 220 survey participants to understand how cognitive biases combined with algorithmic processes influence this process. Our research involves two main approaches: first, an observational study of survey participants studying whether opposing attitudes towards a topic lead to variations in their search queries and subsequently the search results they received; second, controlled experiments using user agents to measure the influence of search history of a user on the search results they are presented with. Our findings revealed two significant insights. First, we observed that while participants with different stances on a partisan issue wrote queries with similar semantic content, their choice of words was significantly different. This suggests that despite differing attitudes, participants were essentially seeking the same information but with implicit variations in vocabulary. Alarmingly, this variation in vocabulary alone was sufficient to skew search results towards results that reinforce their existing beliefs, even when their search history is controlled. Subjects with opposing attitudes were served information from sources that aligned with their beliefs and content that was associated with their preexisting beliefs. We attribute this phenomenon to the collaborative filtering algorithms used by search engines, which appear to construct 'filter bubbles.' These bubbles, shaped by the variations in users' word choices, often present information that supports and potentially amplifies their preexisting beliefs.

How do platforms interpret user inputs? *Ongoing*

Platforms can be presented as simple input and output machines. They are systems that present curated content, for example on home feeds, optimizing for engagement or some other metric for user experience. The input for this system is a library of content and a form of ***user embedding***. This user embedding within the platform represents how the platform perceives the user and is maintained to serve relevant and engaging content to the user. Changes to the perception of the user result in changes in the presented content. Due to the inherent design and black-box nature of the systems we do not know how platforms construct these embeddings (or perceptions). It is then valuable to understand how these perceptions are created and therefore result in the curation of content so that we are well-equipped to understand their outcomes.

In this work, our fundamental goal is to taxonomize platform behavior and create reports for each platform representing characteristics related to how they curate the user's home feed using *user embedding*. We build this understanding of how user embedding is constructed and what it represents by first characterizing ***signals*** afforded by the platform. A signal, defined as a combination of *action* (i.e. like, follow, join, watch, etc.) and *topic* (topic of any kind), is the sole input to the platform that communicates information for the platform to build the user embedding. We seek to standardize these signals across 6 modern platforms (Facebook, X, YouTube, Reddit, Instagram, and TikTok) and compare their influence on the home feeds within the platforms, approximately comparing the processes through which the user embedding is constructed. Investigating similar signals across platforms and statistically comparing their influence on multiple attributes would aid in understanding how the algorithm within each platform *behaves*.

Future Directions

Emergent social algorithmic behavior.

In their review article *Machine Behavior (10),* Rahwan et, al. present platforms as complicated agents and illustrate the need for studying their behavior as they become integral in our society. Platforms and the internet in general rely heavily on algorithms to make decisions. These algorithms range from simple interpretable algorithms to more involved interconnected systems. Platforms often optimize for some metric as they characterize content as its value towards this metric. Therefore, along these metrics, the algorithmic behavior of the platform is somewhat understandable. Because of these metrics, a platform interprets and perceives content completely differently compared to an end user. For example, searching for *U.S. election candidates* on a search engine returns links that optimize for user engagement i.e., they are relevant and of interest to the user. However, platforms have dominated our society as political and social centers—something which they are not measured for. How these systems exhibit emergent social behavior that forms problematic patterns is yet unclear. For example, **why** is it that searching for *U.S. election candidates* yields results that can influence users to change their election voting preferences (11)? The goal is to investigate the relationship of metrics used by algorithms—how platforms perceive content—on the influence of user behavior—how the user perceives the content. This involves understanding the interactions of users with the platforms, how that interaction shapes the platform, and how the platform responds to the users. I am interested in creating methodologies and techniques that can repeatedly and reproducibly measure algorithmic behavior treating them as agents in our society with their influence on different aspects of our lives. I see the goal of these efforts to create a taxonomy of well-studied and documented behavior exhibited by algorithms online over long periods. I am excited to design observational studies such as algorithm audits, experimental studies using sock puppets, and simulations to measure and characterize these behaviors.

User studies.

During my PhD, much of my work has focused on observational studies or experimental studies using user agents (bots). However, even more so than platforms, users exhibit complex behavior that often changes with their background and demographic. To understand the complex relationships between platforms and user behavior, I am interested in performing end-to-end studies that involve user studies in the experiment design. With the combination of computational techniques to analyze platform behavior and social science theories to interpret user behavior, we can present grounded findings that measure the influence of the platform's design decisions on user behavior. In addition to user studies, as they are hard and expensive to scale, I see the limited, yet sufficient user fidelity provided by large language models to be an effective supplement to understanding generalized user behavior. As studies show the effectiveness of LLMs in reproducing the effects of how users consume content to subsequently exhibit their opinions accordingly (12) and interact with other individuals (13, 14), I am excited to utilize them to generalize user studies at large scales. Additionally, to understand how users with different backgrounds experience online platforms, I am interested in understanding how platforms, often built on many Western values, influence users in the global south. With already a lot of valuable and exciting work that investigates these questions (15, 16) from scholars with a diverse range of backgrounds I am excited to pursue understanding how the findings from work focused on the US population translate to users from differing backgrounds. This entails comparing, repeating, and performing original studies across countries and building fundamental generalizable claims about the design of platforms.

References

1. J. Haidt, C. Bail, Social Media and Political Dysfunction: A collaborative review (Unpublished) (January 8, 2024).

2. P. R. Center, Political Polarization in the American Public. *Pew Research Center - U.S. Politics & Policy* (2014) (January 8, 2024).

3. H. Allcott, M. Gentzkow, Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* **31**, 211–236 (2017).

4. , The spread of true and false news online | Science (January 8, 2024).

5. E. Pariser, *The filter bubble: what the Internet is hiding from you* (Penguin Press, 2011).

6. H. Habib, P. Srinivasan, R. Nithyanand, Making a Radical Misogynist: How Online Social Engagement with the Manosphere Influences Traits of Radicalization. *Proc. ACM Hum.-Comput. Interact.* **6**, 1–28 (2022).

7. H. Habib, R. Nithyanand, The Morbid Realities of Social Media: An Investigation into the Narratives Shared by the Deceased Victims of COVID-19. *Proceedings of the International AAAI Conference on Web and Social Media* **17**, 303–314 (2023).

8. H. Habib, M. B. Musa, M. F. Zaffar, R. Nithyanand, Are Proactive Interventions for Reddit Communities Feasible? *Proceedings of the International AAAI Conference on Web and Social Media* **16**, 264–274 (2022).

9. H. Habib, R. Nithyanand, Exploring the Magnitude and Effects of Media Influence on Reddit Moderation. *Proceedings of the International AAAI Conference on Web and Social Media* **16**, 275–286 (2022).

10. I. Rahwan, *et al.*, Machine behaviour. *Nature* **568**, 477–486 (2019).

11. R. Epstein, R. E. Robertson, The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* **112**, E4512–E4521 (2015).

12. E. Chu, J. Andreas, S. Ansolabehere, D. Roy, Language Models Trained on Media Diets Can Predict Public Opinion (2023) https:/doi.org/10.48550/arXiv.2303.16779 (January 26, 2024).

13. J. S. Park, *et al.*, Generative Agents: Interactive Simulacra of Human Behavior in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23., (Association for Computing Machinery, 2023), pp. 1–22.

14. P. Törnberg, D. Valeeva, J. Uitermark, C. Bail, Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms (2023) https:/doi.org/10.48550/arXiv.2310.05984 (January 26, 2024).

15. S. Badrinathan, Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India. *American Political Science Review* **115**, 1325–1341 (2021).

16. I. Danju, Y. Maasoglu, N. Maasoglu, From Autocracy to Democracy: The Impact of Social Media on the Transformation Process in North Africa and Middle East. *Procedia - Social and Behavioral Sciences* **81**, 678–681 (2013).