Overview

Social media platforms have transformed how we consume information. Despite widespread awareness of their significant influence, we have a limited understanding of how platforms shape users' exposure to information and, consequently, their beliefs and behaviors. Within these platforms, various agents and mechanisms—such as curation algorithms, platform governance policies, and the collective actions of users—interact to shape each user's experience. My research examines how these individual mechanisms within platforms contribute to societal outcomes [1-6]. My long-term research goals are to explore platform- and user-driven factors that shape the content curation process and to study alternative mechanisms that promote desirable outcomes for users and society.

Research Focus

I study online platforms oriented towards serving and curating information, such as Google Search, YouTube, Facebook, and Reddit. Due to the nature of their services—content curation—these platforms shape the information users are exposed to, consequently influencing their beliefs and behaviors. While representing platforms as opaque monolithic systems allows us to estimate the net influence they have on users, such representations are limited in measuring causality or establishing the responsible sources of these effects. By studying individual platform aspects, we illustrate how influence is exerted, providing insights beneficial to users and platforms alike. To explore the underlying mechanisms shaping user behavior, my research addresses the following critical questions:

Q1. How do platforms interpret interactions to curate content? In a recent work under review at CSCW [1], I answer this questions by performing large-scale experiments across different platforms, such as Reddit, YouTube, and X. We uncover the configurations of curation algorithms that inform how platforms curate their homepages as they interpret user interactions. For more details, refer to §1.

Q2. How do user preferences and collective user behavior shape problematic personalization patterns? In a mixed-methods study on Google Search [2], we demonstrate how algorithms leveraging collective user behavior yield problematic patterns. We find user's ideological preferences, revealed through their language, combined with learned user behavior from collaborative filtering, reinforced in their search results. Similarly, in another work, we demonstrate, using experimental and observational analysis, the central role collective user behavior plays in shaping recommendations, problematic or otherwise, on YouTube. For more details, refer to §2.

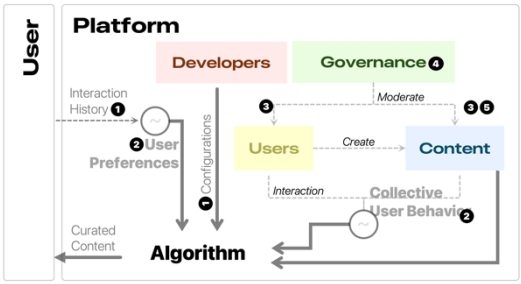

Figure 1 Conceptual model of platform dynamics, illustrating how different mechanisms and agents (users, developers, and governance entities) interact with the platform's algorithm to shape user experience. The numbers in the figure (❶ ❷ ❸ ❹ ❺) correspond to the questions discussed in the text, highlighting the role of user interactions, developer configurations, and governance in influencing how content is curated.

Q3. How do users adopt problematic ideologies? To understand harmful behavior at a user level, in our work published at CSCW [3], we study the adoption of problematic behaviors. Aiming to identify points of intervention, we tracked and illustrated the journey of Reddit users who adopt radical behaviors. Through causal inferences, we demonstrate the influence of key users and dangerous communities in the indoctrination of vulnerable individuals, highlighting their pivotal role in the spread of radicalization—and opportunities for interventions. For more details, refer to §3.

Q4. What drives late community regulation decisions? Platform governance, moderating the user base and content libraries, reduces the harms of platforms. However, when we place platforms within their business contexts, we observe a contrasting reality: administrators often engage in reputation-driven governance. In our observational study on Reddit's community regulation practices [5], published in ICWSM, we discovered that negative media coverage acts as a catalyst for regulating communities on Reddit that violate content policies, indicating that actions are taken more in response to public scrutiny than proactive enforcement. For more details, refer to §3.

Q5. Can problematic community behavior be detected early? Recognizing the challenges faced in large-scale community regulation, in our work published at ICWSM [4], I build a proactive flagging tool for identifying communities evolving towards problematic behaviors, resulting in more complete, informative, and timely moderation. For details, refer to §3.

In the following sections, I elaborate on the key mechanisms shaping user experiences on social media platforms: algorithm configurations (§1), preferences and personalization (§2), and governance (§3). Finally, I outline the future directions of my work (§4).

1. Algorithm Configurations

A significant factor that shapes what content users are exposed to is the algorithms within a platform and how they are configured. For instance, some platforms may prioritize content from a user's network, while others may prioritize recency. These algorithms and their settings are established by platform developers and are designed to optimize user engagement. When curating content for a user, algorithms access conceptual "buckets" of content from different sources or mechanisms—these could include the user's network, contextually relevant content, or personalized content predicted to be engaging to the user. The curated content is shaped by two key factors. First, how these content buckets are filled: for example, what constitutes as "engaging content" can vary between platforms. Second, the proportion of content drawn and ranked from each bucket which shapes the mix of content presented to the user. Together, algorithmic configurations can reveal content exposure patterns.

Platform Interpretation of Interactions.

In a recent work under review at CSCW, I investigate the algorithm configurations of three modern platforms, Reddit, X, and YouTube, to characterize their behavior, uncovering how they interpret user engagements to curate their homepages. By systematically standardizing engagement signals across these platforms, we measure how each engagement signal shapes the home feeds within the platforms. This enables us to approximate the configurations of the algorithms that construct the homepage. To overcome the limitations of limited or non-existent API access, I created sock-puppets that evade sophisticated anti-bot systems while interacting with the platform. This allowed us to isolate the influence of each interaction-topic pair on the homepage of the user across all platforms, consequently generating statistically significant insights into how user interactions shape content curation. We find YouTube's algorithm is highly personalized, driven primarily by *Likes* and passive video consumption, while Reddit's homepage was shaped more by community subscriptions. X displayed the least personalization, with *Follows* having a moderate impact. Our results provide a fundamental understanding of these platforms and the information exposure patterns they can construct, revealing the configurations of their algorithms.

2. Preferences and Personalization

Personalized content optimized for user engagement can shape problematic outcomes. While these outcomes are often attributed to the user, the process of personalization is shaped by three agents: the user, algorithm configuration, and the user base. To curate personalized content, the algorithm first interprets the user's prior interactions to infer their user preferences, a process shaped by the user's actions and the algorithm's configurations. However, to assess the alignment of content with the user preferences, the algorithm leverages the collective user behavior—using techniques such as collaborative filtering. The behavior of the collective user base, therefore, plays a crucial role in personalization. I examine how Google Search and YouTube leverage user preferences and collective user behavior to serve personalized (and problematic) recommendations.

Personalization in information-seeking tasks.

In our work under review at CSCW, we study how Google Search constructs problematic information exposure patterns as a result of collective user behavior and how user preferences are inferred. First, we identify mediums through which the ideological preferences of a user might be revealed to the search engine so that it may present ideologically congruent search results. We find that, apart from the explicit signals present in search history, the language used by users when writing queries betrays their preferences. We study end-to-end information-seeking processes for 220 survey participants to understand how user preferences and 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 receive; second, controlled experiments using user agents to measure the influence of a user's search history on the search results they are presented with. 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 seeking the same information but with subtle 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 leverage collective user behavior to assign relevance to search results and their corresponding search queries.

3. Governance

While platforms present content based on user preferences, collective behaviors, and algorithm configurations, governance serves as an independent mechanism to oversee the content library and user base so that problematic content and harmful

users may be curbed. To this end, platforms use governance teams for moderation, which requires significant labor. Additionally, as the outcomes of social media platforms become complicated and far-reaching, such as radicalization, there is a need for scalable and proactive tools for interventions built on an understanding of how problematic ideologies and behaviors spread.

Community governance.

In a work published in ICWSM, we study how online communities evolve problematic behavior. To understand how ideologies and behaviors infiltrate communities on Reddit, I constructed temporal community embeddings that represent the content and user base of a community for a particular month. By constructing these embeddings, we track how communities evolve. While communities exhibited a general churn and change over time, communities that violated policies showed different patterns even before the violation. Investigating these patterns, we found the insurgence of users from a previously banned community to be the greatest indicator of declining community health. To assist administrators in monitoring community health and curbing the adoption of problematic behavior, we used meaningful features found in our analysis and machine learning techniques to create a proactive flagging tool. In a real-world environment, our flagging tool identified problematic communities months before their ban. This significantly reduces the cost and labor required for moderation, allowing timely moderation—a challenging yet important contributor to effective governance—an insight we explore in another work.

Realizing platforms as businesses with real-world consequences, we examine reasons behind late and incomplete moderation. Through our examinations, we uncover that Reddit's moderations are reputation-driven, where administrative actions on content policy-violating communities can be explained more accurately by negative media coverage than by the underlying violations. By constructing a time series of policy violations (i.e., toxicity) within communities and collecting data on negative media attention, we find evidence of late and inconsistent interventions mediated by negative media attention.

Adoption of radical ideology.

On a more granular scale, in another work published at CSCW, I investigate how users on online platforms, specifically Reddit, develop and exhibit extreme ideologies. We perform a large-scale longitudinal observational analysis on Reddit to make causal inferences on how interaction with dangerous communities and users shapes user behavior. We track and monitor 17,000 users over 68 months, observing changes in their behavior and identifying their interactions with users and communities affiliated with misogynistic ideologies. Using the language within their posts and comments as a window into their behavior and beliefs, we measure the subtle changes as they interact, participate, and get influenced by problematic misogynistic communities. Through a combination of treatment-control and regression analysis, we find interaction with radical users, regardless of where it happens, and the influence of community feedback, positive or negative, key factors in the adoption of problematic ideologies. Through our findings, highlighting the role of problematic communities, we emphasize the importance of timely and complete moderation to prevent the spread of harmful ideologies.

4. Research Agenda

In addition to exploring the mechanisms within online platforms that influence user behavior, my long-term research goals include identifying actionable insights that can reduce the negative impacts of these platforms while preserving their benefits for all stakeholders.

Auditing content curation systems to understand user preferences.

Content curation systems employ different forms of collaborative filtering techniques that analyze and leverage collective user behavior to predict content engagement for a user. These platforms have the advantage of processing and observing massive amounts of interactions between their users and content. These interactions allow the platform—and the algorithms within—to learn patterns in user behavior. Similar to how large language models provide insights into the linguistic behavior of individuals, recommendation and curation algorithms can provide insights into the consumption behaviors of individuals.

However, because we lack direct access to these recommendation and curation algorithms, we must conduct large-scale audits to observe how curated content responds to revealed user preferences. By systematically recording the curated content presented to users based on their preferences, we can construct an approximate mapping between user preferences and the resulting curated content produced by the algorithms. Interrogating this approximation can reveal patterns in user consumption behaviors. More notably, we can identify learned tendencies towards undesirable patterns and uncover user preferences that are vulnerable to problematic outcomes.

By applying this approach across multiple platforms, we can reconstruct learned user consumption behavior patterns in different contexts. This allows us to uncover how affordances across platforms shape user consumption behavior, uncovering

relationships between platform mechanisms and user consumption behaviors. Just as in the early days of computational social science, dense social networks informed how individuals connect online, content curation algorithms, with their rich understanding of social consumption behaviors, will now inform us of how individuals consume content online.

Building a platform test bench.

My prior works and research agenda have so far focused on illustrating the mechanisms within social media platforms that shape user behavior. While it is crucial to identify such mechanisms and their responsibility for problematic outcomes, it is equally important to explore alternative approaches in curating content that might yield improved outcomes. Alternatives that rethink governance or algorithms have the ability to modify the fabric of the platform, inducing emergent effects that require long-term testing in real-world environments. For example, would an algorithm ranking content related to controversial issues in accordance with the Fairness Doctrine decrease affective polarization while maintaining engagement?

Platforms provide limited access for experimental research, making it difficult to test and evaluate alternative approaches. To address the gap in test space for such experiments, I propose the creation of a platform test bench that simulates realistic social media environments. Such a test bench would require faithful reconstruction of internal platform mechanisms that shape users' exposure to information. These include the configuration of algorithms that curate content, a representative collective user behavior that shapes personalization, a content library to curate, platforms governance that regulates user base and content, and realistic user agents with mechanisms to reveal preferences and consume content.

By utilizing the realistic algorithm configurations obtained through my research, we can recreate the processes platforms use to curate content from different sources. Similarly, approximations of curation algorithms can represent collective user behavior based on our prior audits. Understanding individual content consumption behaviors allows us to simulate sets of users with specific decision-making processes, reflecting diverse user groups. By allowing these internal mechanisms to interact within the test bench, we can construct a representative simulation of social media platforms. This framework enables us to meaningfully modify selected designs and mechanisms to measure how outcomes—i.e., curated content and user engagement—are shaped, thus extending our exploration of platform mechanisms into experimental evaluation of potential solutions.

References

[1] H. Habib, R. Stoldt, R. Maragh-Lloyd, B. Ekdale, and R. Nithyanand, "Uncovering the Interaction Equation: Quantifying the Effect of User Interactions on Social Media Homepage Recommendations." arXiv, 09-Jul-2024.

[2] H. Habib, R. Stoldt, A. High, B. Ekdale, A. Peterson, K. Biddle, J. Ssozi, and R. Nithyanand, "Algorithmic amplification of biases on Google Search." arXiv, 17-Jan-2024.

[3] H. Habib, P. Srinivasan, and R. Nithyanand, "Making a Radical Misogynist: How Online Social Engagement with the Manosphere Influences Traits of Radicalization," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW2, pp. 1–28, Nov. 2022.

[4] H. Habib, M. B. Musa, M. F. Zaffar, and R. Nithyanand, "Are Proactive Interventions for Reddit Communities Feasible?," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, pp. 264–274, May 2022.

[5] H. Habib and R. Nithyanand, "Exploring the Magnitude and Effects of Media Influence on Reddit Moderation," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, pp. 275–286, May 2022.

[6] H. Habib and 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*, vol. 17, pp. 303–314, Jun. 2023.