Overview

The information we consume profoundly shapes our beliefs and behaviors. Online platforms have radically transformed how we access information, yet the impact of their complex internal mechanisms—ranging from curation algorithms to governance policies—on users' exposure to information, and consequently on societal outcomes, remains poorly understood. My research aims to uncover how **recommendation algorithms** [1–3], **information flow** [4–6], and **platform governance** [4,7] shape societal outcomes. I use large-scale audit studies, observational studies, machine learning, and natural language processing to diagnose these mechanisms and build strategies and tools to control their harm. My long-term goal is to address *why* the elements within the platform shape user's exposure to information and leverage this knowledge to design interventions that promote positive outcomes. Building on this framework, my current work explores three critical questions:

Q1. How do user preferences shape personalized content curation? To address this question, I first investigate how user interactions are translated into preference signals and then examine how these preferences are reinforced over time—focusing on the magnitude and dynamics of this reinforcement. Through platform audits, controlled studies with simulated accounts, and mixed-methods studies, my work exposes the configurations of these algorithms and the mechanisms by which they learn, reinforce, and amplify user preferences. By identifying patterns in content curation, I explain how algorithmic processes can inadvertently drive phenomena such as affective polarization and filter bubble effects.

Q2. How do ideologies and narratives spread online? An understanding of ideological spread requires examining the phenomenon at both the individual and community levels. I analyze how ideologies are adopted by individuals and employ temporal embeddings to model the behavioral evolution of online communities. Collectively, these studies identify the pathways through which ideologies and narratives propagate online and inform targeted intervention strategies at both the individual and community scales.

Q3. How can platform governance be made more effective? To enhance platform governance, my research identifies systemic weaknesses in current moderation practices and develops scalable, proactive solutions. For example, an investigation into Reddit's community regulations revealed that enforcement is frequently reactive and inconsistent. To address this, I developed a proactive community-health tool that employs discourse and user engagement analysis to identify at-risk communities, thereby enabling timely and effective interventions.

1. Auditing Algorithmic Configurations and Their Outcomes.

A significant portion of content consumed on platforms is curated by content curation algorithms, whose **configurations** play a central role in shaping what users see. Specifically, these algorithms must decide the "source" of curated content—e.g. whether it originates from a user's network or is selected based on user preferences—and determine how to populate this content. I am particularly interested in how preference-aligned content is chosen and how subtle, non-topical signals are incorporated into user preference models.

Uncovering Algorithmic Configurations. To uncover how platforms interpret user interactions to curate their homepages, we conducted an empirical audit [2]—currently under review at FAccT, of the content curation algorithms on three platforms: Reddit, X, and YouTube. We estimated the underlying algorithmic configurations and behaviors by performing standardized interactions and recording the resulting curated feeds. Using a large-scale sock puppet audit, we tested seven interactions (Like, Follow, etc.) across three topics to investigate how user engagements are interpreted to learn preferences. Our results, briefly shown in Fig. 1., reveal the distinct behavioral patterns of each platform.

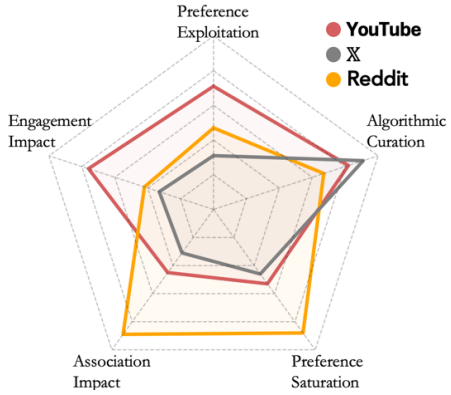

Fig. 1: Algorithmic behavior tendencies of YouTube, X, and Reddit. Preference Exploitation: degree of platform personalization; Algorithmic Curation: personalization by interest-based signals; Preference Saturation: decay rate of repeated preferences; Association Impact: impact of joining or following; and Engagement Impact: impact of like, search, or open.

Measuring Algorithmic Outcomes. In another study [3] under review at CSCW, I demonstrate Google Search's reinforcement of users' ideological preferences by inference of subtle differences in query vocabulary—even when users are seeking similar information. We studied end-to-end information-seeking processes for 220 survey participants which included query formation executed on personal computers. We found that participants with opposing stances on a partisan issue, despite seeking similar information (indicated by semantic similarity within queries), exhibited significant variations in their word choices. The subtle variation alone was sufficient to skew search results towards ideologically congruent results, reinforcing their existing beliefs even under controlled laboratory conditions with cleared search histories. In a recent study [1], we examined whether YouTube's recommendation algorithm can learn and subsequently reinforce emotional preferences. We created "deliberative" sock puppets that revealed their assigned emotional preferences by selecting preference-aligning recommendations.

Our results indicate that subsequent recommendations, even those on topically unrelated content, reinforce and often amplify these preferences.

2. Tracking Adoption of Ideologies and Narratives.

The spread of harmful narratives and ideologies online presents a significant challenge, as evidenced by waves of extremist and conspiratorial narratives spreading from fringe groups to mainstream discourse. My research addresses this by tracking users and communities to diagnose how such content is adopted and to identify opportunities for effective interventions.

Investigating the adoption of problematic ideologies. My research investigates how users on online platforms, particularly Reddit and Facebook, adopt extreme ideologies and share fringe narratives. In work published at CSCW [5], we tracked the adoption of extreme ideologies through a large-scale longitudinal study on Reddit. Analyzing 17,000 users over 68 months, we observed subtle changes in their behavior, measured by the language in user posts and comments, and identified interactions with communities promoting extreme views. Our findings reveal that even subtle shifts in community engagement in user interactions can signal the early adoption of extreme ideologies, underscoring the feasibility and importance of proactive moderation strategies. Building on this work, we further explored the spread of narratives, specifically, pandemic-related misconceptions shared as memes on Facebook [6]. Using computer vision techniques to interpret these memes, we uncovered the evolution of the politicized progression of narratives and the limited completeness of misinformation labels.

Measuring the evolution of online communities. Understanding how online communities evolve and potentially develop problematic behaviors is crucial for fostering healthy online spaces. In our work [4], published at ICWSM, we investigated the evolution of communities in Reddit by developing temporal community embeddings that capture the discourse and user dynamics. While most communities exhibited a general pattern of user and content churn, we discovered that communities that later violated platform policies displayed distinct evolutionary trajectories even before such violations occurred. Leveraging these findings, we developed a community-health monitoring tool to proactively identify communities exhibiting early warning signs of problematic evolution.

3. Improving Platform Governance.

Ideally, platform governance, encompassing user moderation and content policies, should primarily aim to reduce online harms in an effective and timely manner. However, when we place platforms within their business contexts, we observe a contrasting reality: administrators often engage in reactive reputation-driven governance. My research directly addresses this tension by investigating Reddit's governance practices and developing tools for more effective community monitoring.

Investigating reactive governance. Recognizing that platforms operate as profit-driven businesses, we examined the factors behind often delayed content moderation [7]. By constructing a time series of policy violations within communities and collecting data on negative media attention towards them, we find evidence of inconsistent interventions mediated by negative media attention. In other words, administrative actions on content policy-violating communities are more strongly correlated with negative media coverage than with the severity of the policy violations themselves.

Building proactive governance strategies. To empower administrators with more effective and timely governance, we developed a proactive community-flagging tool [4], leveraging key features identified in our community evolution analysis. This tool proactively identifies at-risk communities months before they violate policies. In real-world simulations, our tool demonstrably reduced moderation costs and labor while enabling significantly more timely interventions.

4. Research Agenda

Findings from my work and the broader literature underscore the profound impact online platforms have on society. Traditionally, researchers have focused on measuring either intermediate patterns within platforms—such as filter bubbles, biases, or echo chambers—or the downstream societal outcomes like polarization, extremism, and radicalization. However, these studies often attribute such outcomes to the platform as a whole, without disaggregating the individual components that make up these complex digital systems. In reality, the platform system comprises a constellation of elements—including design decisions, algorithmic configurations, structural affordances, a dynamic user base, and an ever-evolving content library—that interact to produce both intermediate and long-term effects. My long-term research goal is to deconstruct this system and diagnose causal links between its components and the resulting societal outcomes. This goal is operationalized in three interconnected aims: (Aim 1) construct simulations that replicate the system components within the platform and (Aim 2) track narratives in online spaces and assess their evolution as a response to content annotations. Together, these aims will provide critical insights into how targeted manipulations of specific platform components can mitigate harmful outcomes and promote healthier, more balanced digital interactions.

Aim 1: Constructing platform simulations to test platform design and mechanism. This aim bridges the gap between observational studies and experimental analysis. The goal of this research is to address the inherent opacity of these platforms and develop an experimental test bench that replicates their internal mechanisms. In its simplest form, a platform requires a (1) user base that interacts with or contributes to the (2) content library from which (4) content curation algorithms source content to curate on an (4) interface with specific design decisions and affordances. To simulate a platform, we need to replicate and produce these elements.

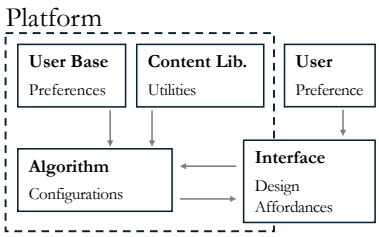

Fig. 2: Structural breakdown of platform dynamics, illustrating how algorithms curate content by interpreting user-revealed preferences through interface interactions. These inferred preferences guide content selection from the platform's library and user base, optimizing for engagement via algorithmic configurations and design affordances.

1) **User base**. Recruiting participants for naturalistic experimental studies provides valuable insights into authentic user behavior and cognition. However, to scale these experiments and explore how platforms respond to variations in user composition, we must also simulate user interactions. Recent advances in large language models (LLMs) have demonstrated an unprecedented ability to learn and reproduce user behavior and personality—effectively mimicking responses from surveys and personality tests [8,9]. In our preliminary work, LLMs trained on users' prior contributions on Reddit successfully captured individual preferences and behaviors. By simulating and manipulating user bases, we can investigate how shifts in collective preferences influence the behavior of other platform elements, particularly how content curation algorithms adapt when presented with an unchanged content library.
2) **Content library.** A thorough understanding of the content library is crucial for investigating how platforms generate problematic exposure patterns—such as amplification effects and filter bubbles—and ultimately influence societal outcomes. We are currently undertaking an experiment that involves scraping and collecting a 1% sample of the vast content library, a representative sample will be sufficient to draw conclusions about the content library.
3) **Content curation algorithms.** Despite often being proprietary, algorithms can be systematically audited to reveal their behaviors and configurations. Building on our previous work in platform auditing, I plan to extend these methods to infer and replicate algorithmic parameters. By reconstructing these configurations and methodically varying them, we can directly assess how subtle changes in algorithm settings affect the content delivered to both real and simulated users.
4) **Interface.** Our research on YouTube's reinforcement of emotional preferences indicates that platforms tend to optimize for implicit, non-deliberative preferences rather than active decision-making. Prior literature, along with our own findings, suggests that frictionless interfaces—characterized by features such as AutoPlay, continuous scrolling feeds, and short-form content—can foster passive consumption patterns that reinforce cognitive biases. By strategically adding or removing interface friction, we aim to study how modifications in user interaction dynamics can shift content curation from bias-driven engagement toward a focus on content quality.

By replicating the fundamental components of online platforms we create a controlled environment to systematically explore the interactions of these elements. This experimental test bench offers a powerful tool to assess how subtle changes in platform design can shape user behavior and content exposure patterns.

Aim 2: Tracking the evolution of narratives in online space to monitor their evolution. Over the past decade, public discourse has shifted noticeably. Distrust in institutions and science has grown as society moves away from a fact-based paradigm toward one dominated by compelling narratives [10,11]. Although researchers have developed a range of design and technical interventions to counter misinformation, a significant portion of the population remains unswayed by these corrective measures. In some cases, attempts to label or fact-check information can inadvertently inoculate and strengthen the narratives they aim to correct [12].

Building on my previous work on how users spread and adopt narratives, this aim focuses on understanding how narratives evolve online over time. Using novel techniques that leverage specialized large language models, I plan to dissect and deconstruct narratives into their fundamental components. This approach will allow us to monitor the transformation of these elements as they interact with exogenous shocks like external events and internal stimuli like misinformation labels.

A central question guiding this research is whether narratives follow a "natural selection" paradigm. Specifically, I will investigate if robust, narratives persist and even strengthen over time, while less resilient narratives—those more easily discredited by factual interventions—fade away. This inquiry not only seeks to map the lifecycle of narratives but also to uncover the mechanisms by which some narratives build an "armor" against corrective efforts. Ultimately, the insights gained from tracking narrative evolution will inform more effective strategies for mitigating the spread of harmful misinformation and fostering a healthier digital public sphere.

References

[1] Habib, H., and Nithyanand, R., "YouTube Recommendations Reinforce Negative Emotions: Auditing Algorithmic Bias with Emotionally-Agentic Sock Puppets." https://doi.org/10.48550/arXiv.2501.15048

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

[3] Habib, H., Stoldt, R., High, A., Ekdale, B., Peterson, A., Biddle, K., Ssozi, J., and Nithyanand, R., "Algorithmic Amplification of Biases on Google Search." https://doi.org/10.48550/arXiv.2401.09044

[4] Habib, H., Musa, M. B., Zaffar, M. F., and Nithyanand, R., "Are Proactive Interventions for Reddit Communities Feasible?," *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16, 2022, pp. 264–274. https://doi.org/10.1609/icwsm.v16i1.19290

[5] Habib, H., Srinivasan, P., and Nithyanand, R., "Making a Radical Misogynist: How Online Social Engagement with the Manosphere Influences Traits of Radicalization," *Proc. ACM Hum.-Comput. Interact.*, Vol. 6, No. CSCW2, 2022, p. 450:1-450:28. https://doi.org/10.1145/3555551

[6] Habib, H., and Nithyanand, R., "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, 2023, pp. 303–314. https://doi.org/10.1609/icwsm.v17i1.22147

[7] Habib, H., and Nithyanand, R., "Exploring the Magnitude and Effects of Media Influence on Reddit Moderation," *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16, 2022, pp. 275–286. https://doi.org/10.1609/icwsm.v16i1.19291

[8] Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S., "Generative Agent Simulations of 1,000 People." https://doi.org/10.48550/arXiv.2411.10109

[9] Chu, E., Andreas, J., Ansolabehere, S., and Roy, D., "Language Models Trained on Media Diets Can Predict Public Opinion." https://doi.org/10.48550/arXiv.2303.16779

[10] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L., "The Science of Fake News," *Science*, Vol. 359, No. 6380, 2018, pp. 1094–1096. https://doi.org/10.1126/science.aao2998

[11] McIntyre, L., "Post-Truth," MIT Press, 2018.

[12] Zuckerman, E., "QAnon and the Emergence of the Unreal," *Journal of Design and Science*, No. 6, 2019. https://doi.org/10.21428/7808da6b.6b8a82b9