Online platforms have become ubiquitous in our lives. Using algorithmic filtering as they curate content optimized for engagement, platforms such as Google, Facebook, and YouTube have fundamentally transformed how we consume information. 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 users adopt behavior and beliefs on platforms and the role different aspects within the platforms such as communities, moderation, and algorithms play in this relationship. My long-term research goals are to investigate how platforms actively influence user behavior and understand how platforms encourage problematic consumption behaviors. Together, these investigations bring insights into the outcomes of platforms, specifically, their harms. In my work, I achieve these goals using two separate frameworks: Observational studies from the user's perspective to uncover how their behavior is influenced and experimental studies to characterize *platform's behavior* to explore how they interpret and respond to user inputs.

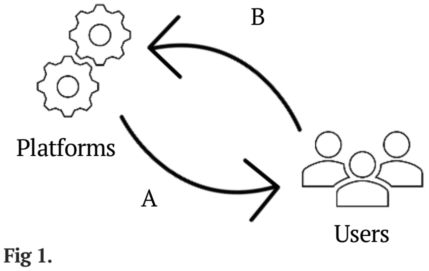

Fig 1.
A. How do platforms influence user behavior?
B. How are user inputs interpreted by platforms?

How do platforms influence us?

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 allows insight into user beliefs and ideologies, crucially, how they changed over time on the platform. To validate relationships between user behavior and platform design, I use statistical and machine learning analyses.

How do users adopt problematic behavior? CSCW 22 • ICWSM 23

I investigate how users on online platforms, like Reddit, develop and exhibit extreme ideologies [6]. Specifically, by monitoring their behavior changes we want to identify engagements that influence the adoption and exhibition of radical misogyny. Using language within user posts and comments as a window to their behavior and beliefs, we measure the subtle changes as they interact, participate, and get influenced by problematic misogynistic communities. We track 17,000 users over 68 months on Reddit to record their interactions and behavioral changes. 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, to be key in the adoption of problematic ideologies. Notably, we found a *recruitment effect* i.e. users who were approached by the radical elites of the manosphere exhibited a significant increase in related radical behaviors compared to their control counterparts. 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.

Further exploring problematic ideologies and their information dynamics, in another study, we explore content shared on Facebook by the victims of COVID-19 who proclaimed their anti-vax beliefs [8]. Using computer vision techniques to interpret and characterize images and memes made by the victims, we analyze the narratives within the posts. We find high politicization of the pandemic and suggestion of epistemic bubbles as users shared from similarly unreliable sources. Both of our works unanimously point towards the overwhelming influence of non-intervened dangerous communities and content, calling first for a need for improved intervention techniques.

How do communities adopt problematic behavior? ICWSM 22 • ICWSM 22

Over the past decade, there has been an insurgence of radical ideologies within online communities. To understand how ideologies and behaviors infiltrate communities, I was interested in exploring how communities *evolve*. To this end, I constructed community embeddings that represent the content and user base of a community for a particular month. Using this on Reddit, we captured the community dynamics and tracked how subreddits (communities on Reddit) evolved [9]. While we found subreddits generally to be extremely dynamic and ever-changing, subreddits that were later banned (due to content policy violations) exhibited significantly different patterns. We investigated this anomaly using feature engineering and machine learning techniques, that yield interpretable and understandable predictions. The features served as meaningful indicators for the decline of communities' health. 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, we leveraged our predictive model to create a proactive flagging tool. We tested our flagging tool in a real-world environment, using continuous learning, where it was able to, on average, proactively flag communities nine months before their actual ban. Our tool allows administrators and the community to identify the declining community health and take proactive steps to address issues. This can significantly reduce the cost and labor required for moderation allowing complete and timely moderation.

Timely moderation is a major challenge community and Reddit faces. To uncover external reasons behind late and incomplete moderation we examine the circumstances surrounding Reddit community bans [10]. We hypothesize that negative media coverage drives interventions. 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. Through our mediation analysis, we discover that content policy violations lead to community bans through their influence on negative media attention. Simply put, Reddit performed reactive community bans driven by negative media attention.

How do platforms interpret user inputs?

Understanding algorithms and internal processes within a platform is challenging due to limited access. Platform outputs such as curated content and recommendation depend on the platform's understanding of user's interests. To understand how the platform interprets user engagements and responds accordingly I conduct experiments using user agents to emulate user behavior and employ a network tomography approach to uncover relationships. This approach allows for characterizing platform behavior, identifying repeated patterns, and detecting potentially problematic tendencies within platforms.

How does Google amplify cognitive biases? *Under review*
Algorithmic retrieval of information has shaped our access to information as platforms like Google become ubiquitous. Yet, the 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 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 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 engagements to curate content? *Ongoing*
Platforms can be considered simple input and output machines. They are systems that curate content optimizing for user engagement or some other metric for user experience. This curation is dependent on how the platforms perceive the user interests. Due to limited access, we do not know how platforms construct these perceptions. In this work, our goal is to characterize platform behavior, uncovering how platforms interpret user engagements to curate content. We define engagement signals as user interactions on platforms, e.g. liking a political post. By systematically standardizing engagement signals across 6 modern platforms (Facebook, X, YouTube, Reddit, Instagram, and TikTok) we measure how each engagement signal shapes the home feeds within the platforms. This enables approximate comparisons of the processes through which the user engagements are interpreted to construct user interests. Investigating similar signals across platforms and statistically comparing their influence on multiple aspects allows us to characterize and compare behavior across platforms and critically identify potential problematic outcomes.

Research Agenda

Content curation is the core service that platforms provide to their users. Through systems, such as recommendation algorithms and search engines, platforms curate content for their users, effectively shaping their exposure to information. An individual's behaviors and beliefs are greatly influenced by their information diet. And, since platforms curate information presented to their users through various algorithms, they have an effective influence on the behaviors and beliefs of their users.

My long-term research agenda is to deconstruct how algorithms within platforms shape users' behavior and beliefs. Through observational and experimental studies driven by cross-disciplinary collaborations, I aim to examine the relationship between platforms and their users. To systematically investigate this phenomenon, my future research goals are to explore the following questions. Contributions from each exploration are crucial in understanding how to control platforms so we may reduce their harm.

What are problematic information exposure patterns?
Our Information diet shapes our beliefs and actions. While the content and its framing play a crucial role in how the information is perceived by the reader, the pattern of exposure of information can be central in shaping how user's perception. I am interested in extending my prior investigations into identifying the patterns of information exposure. Specifically, distortion patterns that shape the information diet, driving the adoption of dysfunctional behaviors, for example, filter bubbles, echo chambers, and epistemic bubbles. Taxonomizing and identifying these patterns within platforms is crucial in identifying how users adopt dysfunctional worldviews. To this end, I am interested in performing end-to-end studies that involve user studies in the experiment design. Through a combination of computational techniques and social science theories, I seek to study when these patterns occur within a platform and how they influence a user. In addition to user studies, I see the limited, yet impressive user fidelity provided by large language models as 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 [13], I am excited to utilize them to support user studies at large scales.

How are problematic information patterns created?
More critically, I am interested in investigating **how** these patterns emerge across platforms i.e., the circumstances under which these patterns emerge. Modern platforms (e.g. Facebook, YouTube, Reddit, Twitter) vary in their affordances and curation mechanisms. In recent work, we design a comparative experimental study to characterize platform behavior and compare the outcomes across platforms. Extending this work, I am interested in systematically identifying the design and algorithmic decisions that yield these patterns. For example, does a recommendation algorithm that optimizes engagement yields an information distortion pattern? To study this I propose large-scale algorithm audits using user agents to uncover these patterns and experimental studies, manipulating the algorithm, to uncover the influence of treatments on the exposure of information.

Why are problematic information patterns created?
Platforms use complex algorithms to curate content for their users. Often algorithms curating content for users are optimized for user engagement. This makes them dependent on user engagement consumption behavior. These algorithms also yield many unwanted emergent effects, for example, affective polarization and widespread misinformation. What causes these effects? Is it driven by user consumption or platform curation?

These emergent effects hinge on the relationship between 1) how users decide to consume content and 2) how platforms decide to recommend content. Users' perception of content and information is driven by complex cognitive processes that include their biases, identities, beliefs, and backgrounds. They interpret various aspects of the content, such as its topic, reliability, and framing, through their personal experiences and understanding. In contrast, platforms' perception of content is largely driven by a more superficial understanding of the material and, more importantly, the content's engagement history—through collaborative filtering. I am interested in investigating how the *perception* of content by platforms results in distorted patterns within content curation. Concretely, uncovering whether platforms *recognize* and amplify problematic consumption patterns—such as cognitive biases. Extending my prior work monitoring how user behavior changes as they consume content, I am interested in developing methodologies that monitor how curated content changes as user behavior shifts. My long-term research goal is to understand the relationship between algorithmic content curation and content consumption driven by users, with a specific focus on user agency within this process. This requires a cross-disciplinary approach, drawing from psychology, sociology, and cognitive sciences. By combining algorithmic analysis with insights into human behavior and societal impact I seek to develop system audits that determine the emergent societal implications of the platform's algorithms.

References

[1] J. Haidt and C. Bail, "Social Media and Political Dysfunction: A collaborative review (Unpublished)." New Yotk University. Accessed: Jan. 08, 2024. [Online]. Available: https://docs.google.com/document/d/1vVAtMCQnz8WVxtSNQev_e1cGmY9rnY96ecYuAj6C548/edit?usp=sharing&usp=embed_facebook

[2] P. R. Center, "Political Polarization in the American Public," Pew Research Center - U.S. Politics & Policy. Accessed: Jan. 08, 2024. [Online]. Available: https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/

[3] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, May 2017, doi: 10.1257/jep.31.2.211.

[4] "The spread of true and false news online | Science." Accessed: Jan. 08, 2024. [Online]. Available: https://www.science.org/doi/full/10.1126/science.aap9559

[5] E. Pariser, The filter bubble: what the Internet is hiding from you. New York: Penguin Press, 2011.

[6] 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, doi: 10.1145/3555551.

[7] T. Grover and G. Mark, "Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit," in *Proceedings of the international AAAI conference on web and social media*, 2019, pp. 193–204.

[8] 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, doi: 10.1609/icwsm.v17i1.22147.

[9] 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, doi: 10.1609/icwsm.v16i1.19290.

[10] 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, doi: 10.1609/icwsm.v16i1.19291.

[11] I. Rahwan *et al.*, "Machine behaviour," *Nature*, vol. 568, no. 7753, pp. 477–486, Apr. 2019, doi: 10.1038/s41586-019-1138-y.

[12] R. Epstein and 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*, vol. 112, no. 33, pp. E4512–E4521, Aug. 2015, doi: 10.1073/pnas.1419828112.

[13] E. Chu, J. Andreas, S. Ansolabehere, and D. Roy, "Language Models Trained on Media Diets Can Predict Public Opinion." arXiv, Mar. 28, 2023. doi: 10.48550/arXiv.2303.16779.

[14] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, in UIST '23. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 1–22. doi: 10.1145/3586183.3606763.

[15] P. Törnberg, D. Valeeva, J. Uitermark, and C. Bail, "Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms." arXiv, Oct. 05, 2023. doi: 10.48550/arXiv.2310.05984.

[16] S. Badrinathan, "Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India," *American Political Science Review*, vol. 115, no. 4, pp. 1325–1341, Nov. 2021, doi: 10.1017/S0003055421000459.

[17] I. Danju, Y. Maasoglu, and 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*, vol. 81, pp. 678–681, Jun. 2013, doi: 10.1016/j.sbspro.2013.06.495.