Hussam Habib | Research Statement

✉ hussam-habib@uiowa.edu • 🖱 www.hussamh10.com

My research seeks to understand the failures of social media design, algorithms and governance in blocking radical and manipulative pathways. By uncovering the process of how users adopt extreme opinions online and the role social media platforms, more specifically the online communities, algorithms, and moderation, play in this process, the long term goal of my research is to improve platform design and implement intervention strategies as to not enable these pathways. In my research projects, I use a combination of social psychology theories, natural language processing and statistical analysis to identify the complex social processes online.

In recent years, social media platforms have played a key role in shaping our socio-political climate. Numerous incidents of violent outbursts by radicalized individuals from fringe online communities have made the substantial impact of social media platforms on our beliefs, actions and ideologies visible.

To develop safe and welcoming communities, we first need to develop an understanding of the social processes in current platforms. *My current projects aim to answer 1) how online communities and actors facilitate adoption of extreme and problematic opinions and 2) whether the current moderation strategies are effective in preventing this process.*

Reactive community-level interventions

This project [**?**] currently under review at ICWSM 2021, seeks to understand the current community-level moderation decisions taken by Reddit. The goal of the project is to measure the influence of the media's negative attention on Reddit's administrative decisions. Using statistical mediation analysis on previous moderation decisions and articles from popular media outlets we provide evidence of a reactionary pressure-driven administrative strategy employed by Reddit for closing toxic communities. Our results establish Reddit's reliance on reports from media outlets for closing dangerous communities and in doing so highlight the inconsistencies in Reddit's moderation policies and their enforcement. Next, we seek to evaluate the drawbacks of reliance on media reports for interventions specifically by testing whether media reports exacerbate the problem worse by inadvertently popularizing dangerous ideologies. Using natural language processing we track and measure the growth and spread of dangerous ideologies and communities. Our interrupted time series analysis shows an increase in the adoption of some dangerous ideologies after the initial media reports and interventions. This suggests media's criticism can inadvertently lead to the popularization of the ideology and in some cases can make interventions ineffective. Taken together, our results provide evidence for Reddit's reactionary moderation to maintain a positive reputation in media for advertisers. Our results highlight the need to reevaluate Section 230 and whether platforms, using current moderation strategies, can moderate themselves effectively.

Feasibility of pro-active moderation

In this project [**?**], published in ICWSM 2022, we aim to test the feasibility of pro-active moderation of dangerous communities. First, We study the topical and participant evolution of online communities on Reddit. Using text analysis and statistical methods we construct and track community embeddings to determine their evolution. Our results show communities as non-static and constantly evolving structures requiring constant human moderation. Furthermore, by tracking the evolution of communities we uncover the distinct evolutionary patterns of dangerous communities. Our results show the evolution of a benign community towards toxicity is significantly more unstable. Given the non-static nature of online communities, the need for prohibitively expensive constant human moderation and the distinct evolutionary pattern for dangerous communities, we test the feasibility of proactive flagging of devolution of communities for moderators further consideration. We use interpretable machine learning techniques on features gathered from the structure, user-base, content and external attention of the banned communities to train a classifier to classify the devolution of communities. Comparing and evaluating our tool with Reddit's community-level administration we demonstrate high accuracy of our tool along with inconsistencies in Reddit's enforcement of its policies. In conclusion, this work validates the feasibility of proactive assistive flagging of dangerous communities to alleviate the need for constant human moderation.

Anti-feminist radical pathways

In this project, we seek to identify the events which lead to the online radicalization of individuals. This project focuses on the radicalization of users on the topic of anti-feminism. We observe the effects of online communities, actors, and experiences on user behavior and seek to identify any radicalizing events. Using a combination of social psychology theories and text analysis we measure and track the warning behaviors associated with the radicalization of individuals online along with any event they experience. Next, using statistical analysis and machine learning algorithms, we identify radicalizing events that cause an increase in their warning behavior. We define these events as radical pathways towards anti-feminism and explore the level to which each event radicalizes the individual. The goal of the project is to identify the role online actors play in radicalization. The findings of this project lead to the question of how much do the platform design and governance enable these radicalizing online actors. In the next project, we aim to answer this question and propose intervention strategies in blocking these pathways.

Future Projects

My future projects focus on the role of (mis)information and personalization algorithms in manipulation of user beliefs. While my current work focuses on online actors, specifically, online communities and platform moderators, in their role in the process of manipulation of individuals beliefs and the adoption of problematic and extreme ideologies, my future projects would aim to understand and highlight the role of information and algorithms in this process as non-actors. My future projects are designed to answer how information spreads in communities, how individuals seek information online and how algorithms are exploited to spread information. The eventual goal of my research is to help design better online experiences and mitigate the negative consequences of online communication. By growing the understanding of processes involved in online radicalization and polarization, and highlighting the role of platform design, governance, information, and online communities my research aims to provide intervention strategies and design decisions to aid the construction of healthier online communities.