The Groupthink Bias: Algorithmic Echoes Across Social Media

Junior Francisco Garcia Ayala

Department of Psychology, New York University

PSYC-GA 2083 Group Dynamics

Dr. Azadeh Aalai

December 12, 2023

The Groupthink Bias: Algorithmic Echoes Across Social Media

Introduction

Social media has become the modus operandi of extremist groups to further their recruitment, mobilization, and voice amplification processes, all of which can culminate in disastrous events like the January 6th insurrection or the Rohingya genocide in Myanmar. These cult-like insurgencies have been orchestrated entirely through social media platforms in a distributed manner, amplifying the voices of extremist groups/actors that were once at the peripheries of public discourse. For example, it is reported that 1,106 defendants, all civilians who were called to arms through social media platforms like Twitter, were charged for actions they took during the January 6th attack on Capitol Hill (Sforza, 2023). Additionally, Amnesty International has reported that Facebook's algorithm is to blame for promoting anti-Rohingya content, a movement which has displaced over 700,000 Rohingya Muslims to nearby countries (De Guzman, 2022).

Although anti-social behavior in online communities has been rigorously proven to be a problem in various peer-reviewed studies (Park et al., 2022), the specific group mechanisms predisposing individuals to extremist tendencies warrant deeper exploration. This paper will explore the psychological underpinnings of group dynamics within and enabled by social media platforms. Drawing from theories such as groupthink (Tsintsadze-Maass, 2022), identity formation (Kampmier, 2021), and cognitive convergence (Parunak et al., 2008), the factors influencing users to conform to dominant opinions, often suppressing or altering their beliefs, will be examined.

Background

Human beings are inherently social creatures, and it is worthwhile to ground this work in the socio-theoretical underpinnings that inform our understanding of human sociability. In the psychology of the self, the self-concept is defined as the collection of beliefs one holds about oneself, typically in response to the question, "Who am I?" (Gecas, 1982). Social identity theory extends this notion by championing the idea that one's self-concept is also partly due to one's membership in a social group (Harwood, 2020). Indeed, the need to belong to a group is so fundamental that it can be considered a quasi-physiological need, as evidenced by Maslow's hierarchy of needs, which ranks "Love and Belonging" as the third most crucial category, just after "Safety and Security" and "Physiological Needs" (Mcleod, 2022). The interplay between personal and group identity becomes particularly evident in the dynamics of in-groups and out-groups. An individual's sense of self is affirmed through their in-group membership and the differentiation from and often alienation of out-groups. This dichotomy of 'us' versus 'them' solidifies one's identity within their group and serves as a psychological mechanism that reinforces allegiance to the group, further intertwining the individual's self-concept with their social identity (Tajfel, 1974).

While the original intention behind their creation remains debatable, technologists and engineers created social media platforms that now act as a physically unbounded extension of daily life. The benefits platforms like Facebook, Twitter, YouTube, Instagram, and Reddit provide are undoubtedly manifold. They offer ubiquitous connectivity, allowing individuals to maintain relationships over vast distances, share experiences in real-time, and access a wealth of information at their fingertips. These platforms have also increased the ability by which regular individuals can create and disseminate content, giving voice to those who might otherwise

remain unheard in the traditional media landscape. These benefits have come at the cost of the free-reign the creators behind these social media platforms have to manage the algorithms that power the content feeds their users consume, prompting governments and international organizations to devise plans and legislation to keep these platforms accountable (O'Hagan, 2023). Although some work has been done to perform exhaustive content audits by treating these algorithms as a black box that returns an output followed by a specific set of inputs and systematically evaluating the effects of these outputs (Ibrahim et al., 2023; Juneja et al., 2023), the fact remains that what happens inside said black-box remains largely unknown, despite some recent premature efforts in open-sourcing them (Twitter Team, 2023).

Social Media and its role in radicalization

The role of social media algorithms in radicalization and deepening societal divides has come up quite frequently in the literature. For example, Juneja et al. did a posthoc analysis of YouTube's recommendation algorithm to understand how the platform's algorithm surfaces election misinformation. Given the relevance platforms like YouTube have today on the regular user's information landscape, it is essential to maintain the platforms accountable for the content they serve their user base, especially if such content contains misinformation that can lead to radicalization, which usually leads to more considerable dangers to society at large, like the Capitol Attack. The authors conducted their study by meticulously selecting search queries and videos concerning the Presidential Election and claims of voter fraud in the 2020 elections. These were used as initial content for 99 participants who viewed YouTube through a special extension called TubeCapture, created by the authors. This extension recorded the video recommendations generated after the seed queries and videos were entered into the participants' YouTube browsers under controlled conditions. The videos collected by TubeCapture were then

examined to evaluate the extent of misinformation they contained, using a blend of expert reviews and machine learning techniques. This data was used to determine how the platform deviates users away from (or into) election misinformation. The authors found that YouTube pays more attention to queries with misinformation leanings and ensures users are exposed to debunking videos. Although misinformation videos are well reduced in the platform, there is still room for improvement in the up-next trail of recommendations. This might be because the recommendations in this trial are content-based instead of user-based, so if a user is watching a misinformation video, they might get a misinformation video recommendation (Juneja et al., 2023).

Another work that studied the prevalence of anti-social behaviours in online communities is Joon Sun Park's "Measuring the Prevalence of Anti-Social Behavior in Online Communities" (Park et al., 2022). The authors recognize in this work that empirical studies of anti-social behaviours in online communities, those enabled by social media platforms, are deeply understudied, "with platform-published performance metrics often tucked under the platforms transparency or compliance reports using vaguely defined categories". The few studies that have engaged in this type of work have demonstrated that such anti-social behavior, if visible and widespread, can encourage other people to participate in it as well, a stance also adopted in this analysis (Cheng et al., 2017). Using a novel human-AI pipeline, the authors found that 6.25% (95% Confidence Interval [5.36%, 7.13%]) of all comments in 2016, and 4.28% (95% CI [2.50%, 6.26%]) in 2020-2021, are violations of a set of macro-norm violations, established by a prior work, in the 100 most popular Reddit subreddits. This work is particularly relevant because even in platforms that are primarily self-governed, toxic behaviour remains largely prevalent. Moreover, the moderators of these self-governed platforms do so without any monetary

compensation and are largely overworked. The identification of these norm violations is important in providing a proof-of-concept system that could eventually empower moderators to cover all possible norm violations more expansively in real time. However, developing algorithms that can do this in real-time remains a challenge. Furthermore, given the bureaucratic nature of these social media platforms, embedding crowd-sourced moderation carried out by an impartial third party remains an open question. How much agency should these platforms give to the moderators to perform their work? Juneja et. al.'s work showed that dangerous misinformation content remains largely available on YouTube in the up-next trial recommendations of a video, which signifies that their algorithmic counter-misinformation work remains largely ineffective and clearly unmonitored, begging the necessity for more impartial oversight.

It is also important to adopt a devil's advocate perspective and praise the efforts done by social media platforms in embedding societal values into their algorithms, even if such efforts come from the pressure of larger institutional and governmental demands inspired by the aftermath of dangerous societal harms caused. For example, Mark Zuckerberg, the CEO of Meta, underwent serious scrutiny by the U.S. Congress when they asked him over 600 questions regarding Facebook's mishandling of user data, as evidenced by the infamous Cambridge Analytica scandal (Wichter, 2018). In doing so, we can learn from these efforts, even if they are still in their infancy, the necessary engineering and societal work needed to design better algorithms. This is a stance adopted by Bernstein et al. in their commentary work on embedding societal values into social media algorithms (Bernstein et al., 2023). The authors note how Facebook built models that could predict and subsequently downrank posts that people dislike even if they are likely to click on them to reduce overall clickbait content, how some platforms

weigh the effect of user feedback on other users who might otherwise get few replies, and that most well-established social media platforms employ entire trust and safety teams that collaborate directly with the algorithms to flag and remove content following a bridge of a set of community standards. Some work has even received feedback from a battery of these trust and safety teams to develop cyber-security-inspired frameworks to counter misinformation (Mirza et al., 2023). One example of a platform openly sharing the inner workings of its recommendation algorithm is Twitter, now X, which employs an AI model that synthesizes signals such as content, type of posts, and additional models trained on satisfaction surveys to create predictions on the probability of " (1) the user will reply to the post, (2) the post's original author will engage with that reply, (3) the user will engage with the author's profile page, (4) the user will retweet and share the content, (5) the user will give negative feedback such as not interested in this post, and others (Bernstein et al., 2023)." Hyper-optimizing for user-specific engagement, even in the presence of AI-powered filters that might remove some harmful content, can become a double-edged sword. The authors note, for instance, that, in the US, users are more likely to post positive comments but are more likely to be influenced by negative content and subsequently share the content with others as it might violate their values and hijack their attention. As such, a natural next step would be for the algorithm designers to lift their gaze and prioritize larger societal values as opposed to more user-focused engagement. Tractably operationalizing such societal values requires a translation of abstract ideas like equality, safety, and misinformation into mathematical objective functions that can be understood by algorithms. Although the authors cite interesting work they did that translates democratic values into objective functions leveraging large-language models like Chatgpt (Jia et al., 2023), establishing who determines what a societal value is remains an open problem. For example, in the United States context, is

removing disinformation from social media platforms a democratic or anti-democratic move? The ethos of American society believes that the costs of unilaterally allowing a platform to remove such hateful content outweigh the costs of allowing such content to remain in the first place as a legacy of the First Amendment in the US Constitution. Moreover, these platform-level definitions of societal values might hurt marginalized groups since, in practice, these groups do not have equal access to public spheres enabled by social media platforms, as argued by Fraser (Fraser, 1990). As such, a better approach is to allow some leeway in how these groups can come into existence and study the dynamics of group formation in this context to find ways to resolve such naturally occurring value conflicts.

Dissecting the groupthink phenomenon

In this paper, we adopt the view that group identity, which is part of the basis for self-identity, is amplified by social media platforms. From a technical perspective, it is evident that content recommendation algorithms induce a group cascade effect, whereby one's social network influences the content viewed in conjunction with other factors, such as metadata associated with the user's previous interactions with content. Once the content appears within a user's view, the user's propensity to engage with it is already heightened as it is a mere reflection of the communities a user is a part of (Medvedev, 2019). This engagement can be manifested in many ways, and the one that is of most interest is the comment section of a post, as it signals a higher level of engagement when compared to a like or a repost. This can be attributed to the "online disinhibition effect", where users, protected by a guise of anonymity, act out more frequently and harshly than they would in person (Suler, 2004). Not only does this empower users to act, but they often do so negatively due to the lack of consequences their comments incur. The comment section, therefore, warrants further investigation as a potential hotbed for

groupthink.

Groupthink, a term introduced in 1972 by psychologist Irving Janis, is described as "the mode of thinking that individuals engage in when the pursuit of agreement becomes so dominant within a cohesive in-group that it tends to override the realistic appraisal of alternative courses of action" (Tsintsadze-Maass & Maass, 2014). According to Janis, groupthink is based on five antecedent conditions: high levels of stress due to external threats, the insulation of the group from outside perspectives, homogeneity of members' social background and ideology, no tradition of impartial leadership, and a lack of codified decision-making procedures. In the context of terrorist radicalization, groupthink can compel individuals to engage in terrorist behaviour, even when such actions are futile. Initial conversion to and involvement with a terrorist organization can be attributed to the amplification of regular psychological processes. One example can be the grievance caused by an impactful event that motivates a rational radicalization process where violence is seen as the only way to effect change that seems impossible through peaceful means. Another example could be the in-group versus out-group dynamics that signal an undying commitment to a cause, further distinguishing the group from its adversaries and attracting new members to its ranks. Regardless of the reason, once the allegiance is established, persistently engaging in a cycle of terrorist activities with no end in sight can be seen as a form of cognitive convergence toward what the group desires collectively.

Tsintsadze-Maass & Maass illustrate the phenomenon of groupthink through a case-study deep dive into a terrorist organization of the 1960s named Weather Underground. Weather Underground was born out of the iconic youth activist organizations of the New Left, a counterculture movement of the 60s which emphasized leftist ideals with a focus on realizing social justice and "an emphasis on cultural as well as political transformation, an extension of the

traditional left's focus on the class struggle to acknowledge multiple forms and bases of oppression, including race and gender, and a rejection of bureaucracy and traditional forms of political organization in favour of direct action and participatory democracy" (Davis, 2023). Their primary goal was "the creation of a mass revolutionary movement to overthrow the U.S. government, which it considered the main source of atrocities worldwide" (Tsintsadze-Maass & Maass, 2014). Their analysis was based on breaking down the ways this group fits Janis' antecedent conditions for groupthink. The group experienced *high stress from an external threat* because it came into existence during an era of heightened political violence, as evidenced by the millions of youths being conscripted into the military to fight the Vietnam War, and the assassinations of Martin Luther King Jr., Robert Kennedy, and Black Panther leader Fred Hampton. *The group's insulation from outside influences,* further minimizing exposure to alternative perspectives, came when the group isolated itself from its parent group, Students for a Democratic Society, due to a radical and gradual adoption of Marxist ideology and militancy, contrasting with the SDS's milder attempts to collaborate and negotiate with politicians. *Homogeneity of the group's social background and ideology* occurred as most of the group's members were young, white, former students from affluent families. *A lack of impartial leadership*, magnified by a preference to employ over-the-top rhetoric that motivates emotional loyalty from members, occurred in Weather Underground through a cadre of leaders who embodied such a personality by 'refusing to let anyone else talk until they finished, and their tone was so commanding and contentious that everybody in the room was more or less intimidated into silence.' Finally, the group partook in *defective decision-making processes* by having an incomplete, often blurry, set of objectives, diminishing alternative policies, focusing on biased and selective information, and negating possible contingencies despite failure.

Conclusion

Similar to how Tsintsadze-Maass & Maass were able to characterize how Weather Underground met the five antecedent groupthink conditions, a similar line of work is needed to characterize in real-time if online communities are meeting these conditions. In doing so, we could begin the de-radicalization process pre-emptively to disable these groups from becoming a threat to society at large. For example, some relevant processes necessary to begin the de-radicalization process include improving social surroundings, experiencing acceptance and inclusion, improving personal status and self-image, finding structure and stability, and disillusionment and maturing out (Reiter et al., 2021). Interestingly enough, Reiter et al. also noted how there are parallels between the radicalization and de-radicalization processes: how "compensating for domestic and social problems during radicalization is paralleled by improving social surroundings during deradicalization". If we leverage the parallels between the radicalization and de-radicalization processes and are able to elucidate the intricacies of these processes to form tractable value functions following the recommendations of Bernstein et al., we could clearly show the logical gaps that exist within an online group's values and provide recommendations both to the group itself, the moderators of the online communities, and any interested third parties. This can follow from a process of natural language data extraction and analysis from social media platforms, similar to what Park et al. did in their analysis of anti-social behaviour in online communities. For example, such a process could analyze if two opposing Reddit online communities, r/democrats/ and r/Republican/, are satisfying Janis' groupthink antecedent conditions. We could then compare the analysis between both groups and see if there could be similarities in their radicalization process to prescribe the appropriate deradicalization responses.

References

Bernstein, M., Christin, A., Hancock, J., Hashimoto, T., Jia, C., Lam, M., Meister, N., Persily, N., Piccardi, T., Saveski, M., Tsai, J., Ugander, J., & Xu, C. (2023). Embedding Societal Values into Social Media Algorithms. *Journal of Online Trust and Safety*, *2*(1). https://doi.org/10.54501/jots.v2i1.148

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1217–1230. https://doi.org/10.1145/2998181.2998213

Davis, M. (2023, December 12). *New Left | Definition, History, & Facts | Britannica*. https://www.britannica.com/topic/New-Left

De Guzman, C. (2022, September 28). *Report: Facebook Algorithms Promoted Anti-Rohingya Violence | Time*. https://time.com/6217730/myanmar-meta-rohingya-facebook/

Fraser, N. (1990). Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. *Social Text*, *25/26*, 56–80. https://doi.org/10.2307/466240

Gecas, V. (1982). The Self-Concept. *Annual Review of Sociology*, *8*(1), 1–33. https://doi.org/10.1146/annurev.so.08.080182.000245

Harwood, J. (2020). *Social Identity Theory*. 1–7. https://doi.org/10.1002/9781119011071.iemp0153

Ibrahim, H., Aldahoul, N., Lee, S., Rahwan, T., & Zaki, Y. (2023). Youtube's recommendation algorithm is left-leaning in the united states. *PNAS Nexus*. https://doi.org/10.1093/pnasnexus/pgad264

Jia, C., Lam, M. S., Mai, M. C., Hancock, J., & Bernstein, M. S. (2023). *Embedding Democratic*

Values into Social Media AIs via Societal Objective Functions (arXiv:2307.13912). arXiv. https://doi.org/10.48550/arXiv.2307.13912

Juneja, P., Bhuiyan, M. M., & Mitra, T. (2023). Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on YouTube. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–22. https://doi.org/10.1145/3544548.3580846

Mcleod, S. (2022, November 3). *Maslow's Hierarchy of Needs*. https://www.simplypsychology.org/maslow.html

Medvedev, I. (2019, November 25). *Powered by AI: Instagram's Explore recommender system*. https://ai.meta.com/blog/powered-by-ai-instagrams-explore-recommender-system/

Mirza, S., Begum, L., Niu, L., Pardo, S., Abouzied, A., Papotti, P., & Pöpper, C. (2023). Tactics, Threats & Targets: Modeling Disinformation and its Mitigation. *Proceedings 2023 Network and Distributed System Security Symposium*. Network and Distributed System Security Symposium, San Diego, CA, USA. https://doi.org/10.14722/ndss.2023.23657

O'Hagan, C. (2023, November 6). *Online disinformation: UNESCO unveils action plan to regulate social media platforms | Articles*. https://www.unesco.org/en/articles/online-disinformation-unesco-unveils-action-plan-regulate-social-media-platforms

Park, J. S., Seering, J., & Bernstein, M. S. (2022). *Measuring the Prevalence of Anti-Social Behavior in Online Communities* (arXiv:2208.13094). arXiv. http://arxiv.org/abs/2208.13094

Parunak, H. V., Belding, T. C., Hilscher, R., & Brueckner, S. (2008). Modeling and managing collective cognitive convergence. *Proceedings of the 7th International Joint Conference*

on Autonomous Agents and Multiagent Systems - Volume 3, 1505–1508.

Reiter, J., Doosje, B., & Feddes, A. R. (2021). Radicalization and deradicalization: A qualitative analysis of parallels in relevant risk factors and trigger factors. *Peace and Conflict: Journal of Peace Psychology*, *27*(2), 268–283. https://doi.org/10.1037/pac0000493

Sforza, L. (2023, August 10). Number of people charged in Jan. 6 rioting surpasses 1,100 [Text]. *The Hill*. https://thehill.com/policy/national-security/4147038-number-of-people-charged-in-jan-6-rioting-surpasses-1100/

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information*, *13*(2), 65–93. https://doi.org/10.1177/053901847401300204

Tsintsadze-Maass, E., & Maass, R. W. (2014). Groupthink and terrorist radicalization. *Terrorism and Political Violence*, *26*(5), 735–758. https://doi.org/10.1080/09546553.2013.805094

Twitter Team. (2023, March 31). *Twitter's Recommendation Algorithm*. https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm

Wichter, Z. (2018, April 12). 2 Days, 10 Hours, 600 Questions: What Happened When Mark Zuckerberg Went to Washington. *The New York Times*. https://www.nytimes.com/2018/04/12/technology/mark-zuckerberg-testimony.html