Steve Rathje – Research Statement

Humans are a social species, yet our social environment has been changing rapidly. While we evolved to cooperate and share information in small groups, messages—amplified by opaque algorithms—are now rapidly spreading on social media to millions of strangers around the globe. What makes information (or misinformation) go "viral" or spread widely online? Are the kinds of messages that spread offline different from the ones that spread online? How does exposure to (mis)information online and offline affect important psychological outcomes, such as intergroup conflict, polarization, and well-being?

Much of my research explores the psychology of "virality," or the science of why information spreads in both online and offline contexts. While this question is relevant with the rise of social media, messages can also go "viral" and spread widely offline, and the offline spread of rumors has been explored in classic social psychological work (Allport, 1947).

To study this topic, I take a multi-method approach, employing computational social science techniques, lab experiments, longitudinal field experiments, and multi-site global studies. I ground this empirical work in classic social psychological theories, such as social identity theory (Tajfel & Turner, 1979), and motivated reasoning (Kunda, 1990). I expand on these theories in the context of the digital age, exploring how emerging technologies such as social media and artificial intelligence interact with intergroup conflict and the spread of (mis)information. I also explore the long-term consequences of (mis)information exposure, and how the spread and impact of (mis)information differs in different contexts around the globe.

The Psychology of Virality

To understand what goes "viral" online, my colleagues and I analyzed nearly 3 million social media posts from congress members and partisan news media sources on Facebook and Twitter. While most prior work has examined how emotions (such as negative emotions, high-arousal emotions, or moral emotions) predict virality, we drew on social identity theory (Tajfel & Turner, 1979) to examine how expressions of in-group favoritism and out-group derogation shape virality in a political context.

We found that social media posts about the political out-group – but not the political in-group – were very likely to go viral. Out-group language was a much stronger predictor of virality than established predictors, such as moral or emotional language (Rathje, Van Bavel, & van der Linden, *PNAS*, 2021). Specifically, each additional word about the political outgroup added to a social media post led to a 67% increase in the number of shares that post received. Further, words about one's outgroup were strongly predictive of "angry" and "haha" reactions, likely indicating outgroup derogation. These results, which have since been replicated multiple times (Yu et al., 2023; Heltzel, 2024), suggest that people, politicians, and media companies have perverse incentives to post polarizing content about their out-groups on social media, since this is the type of information that captures people's attention and goes "viral" (Rathje & van der Linden, *Research Handbook on Nudges and Society*, 2023; Van Bavel… Rathje, S., *Annual Review of Psychology*, 2023).

The Paradox of Virality

People engage with polarizing content online, but does this mean that people *like* polarizing content and want it to go viral? In a recent paper, we asked a nationally representative sample of United States participants what they think goes viral versus what they *want* to go viral on social media (Rathje, Robertson, Brady, & Van Bavel, *Perspectives on Psychological Science*, 2023). Most people reported thinking that divisive content, misinformation and moral outrage go viral online, suggesting that people have some awareness that divisive content is amplified in online social networks. However, the vast majority of our sample – Republicans and Democrats alike – reported that they *do not want* this type of content to go viral. This introduces a paradox: even though people tend to *engage* more with divisive content online, they report not liking this content and not wanting it to go viral. In other words, there is a large gap between people's online behavior and their stated preferences.

Shifting Incentives

Why do people paradoxically engage with information that they say they do not like? One potential reason might be the incentive structure of social media, which draws our attention toward partisan identity motivations and diverts our attention from accuracy motivations. To test this hypothesis, my colleagues and I conducted a series of four online experiments with more than 3,000 participants (Rathje, Roozenbeek, Van Bavel & van der Linden, *Nature Human Behavior*, 2023). We found that motivating people to be accurate (through financial and social incentives) made people more accurate at discerning between true and false news and reduced partisan bias in judgements of news. Conversely, incentivizing people to identify news that would be liked by their political allies *decreased* people's accuracy at discerning between true and false news headlines. These findings suggest that social media's incentive structure—which is primarily about rewarding people for engagement—may interfere with accuracy. They also suggest that interventions that incentivize accuracy and disincentivize partisan signaling can improve online and offline environments.

Online Social Networks

In addition to exploring how (mis)information spreads, I explore the consequences of exposure to (mis)information in one's online social network. To do this, I have linked social media data to survey data to explore how the structure of one's online social network is associated with offline attitudes. For example, in one study, my colleagues and I found that following, favoriting, or retweeting low-quality news sources was associated with vaccine hesitancy – even when controlling for other variables, such as self-reported ideology or education (Rathje, He, Roozenbeek, Van Bavel & van der Linden, 2023, *PNAS Nexus*). We also conducted network analysis, which revealed that vaccine-hesitant and vaccine-confident individuals were in distinct digital "echo chambers" in the US and the UK. Further, centrality in a conservative-leaning online network in the US (but not the UK) predicted vaccine hesitancy – illustrating that vaccine attitudes became tied to partisan identity in some cultural contexts.

Experimentally Manipulating These Networks

While the above work helps identify a link between online social networks and offline attitudes, it is correlational and does not test the *causal* impact of exposure to false or polarizing information. To test the causal impact of exposure to (mis)information in one's online network, my colleagues and I conducted multiple large-scale, longitudinal field experiments, funded in part by a $175,000 grant I received from the Russell Sage Foundation, in which we incentivized more than 1,600 social media users to *unfollow* "polarizing" influencers and low-quality news accounts for one month, and then tracked participants attitudes and online behaviors for a full year.

Unfollowing "polarizing" accounts reduced out-party animosity, with effects lasting up to six months. Only 42% of participants chose to refollow polarizing accounts when the experiment was over, and for the subset of participants who did not refollow accounts, this effect on out-party animosity lasted a *full year*. Unfollowing also changed behavior, making people like and repost more accurate news accounts. Furthermore, it made people feel better about their

Twitter/X feeds, and report seeing less political content one year later—without reducing online engagement.

Unlike other interventions that ask people to simply reduce their social media usage, this is a targeted intervention: like a scalpel, it surgically removes a few carefully selected harmful parts of one's feed, allowing the positive aspects of social media to remain. And unlike most behavioral interventions, which often have fleeting effects, this was a structural intervention that changed the daily content of one's online information diet, which may be why it had such lasting effects on beliefs and behavior. Building on my prior "big data" work showing that out-group animosity goes viral, this study shows that frequent exposure to polarizing content from social media influencers in one's daily information diet has a lasting, causal impact on beliefs and behavior.

Exploring These Questions on a Global Scale

My future work aims to test these questions on a global scale. While most social media research focuses on the United States, there is good reason to believe that the psychological impact of social media on intergroup conflict and mental health differs in different contexts around the world. For instance, meta-analyses suggest that the impact of social media usage might be very different in less Democratic nations (Lorenz-Spreen et al., 2021) or in the Global South (Ghai et al., 2022). A full theoretical account of how social media shapes polarization, intergroup conflict, and other variables will require more global and causal evidence (Van Bavel, Rathje, Harris, Roberson, Sternisko, 2021, *Trends in Cognitive Science*; Harris, Rathje, Robertson & Van Bavel, 2023, *International Journal of Communication*).

I am currently leading a multi-site, global field experiment that is testing the causal impact of reducing social media usage on several psychological variables in dozens of countries around the globe. I led the writing of several large grants for this project and received more than $1.6 million in grant funding from several agencies (including the National Science Foundation) to conduct this experiment. I am leading more than 600 collaborators from over 70 countries to help collect data and translate surveys for this multi-country experiment. We have been invited to submit this project as a Registered Report at the journal *Nature*. This large-scale experiment will help answer hotly debated questions about the causal impact of social media usage on polarization, intergroup conflict, and well-being around the globe. It will also allow us to explore how information from people's online social networks interacts with information from their offline social networks and broader cultural context.

Network-Level Effects

While experiments that encourage participants to reduce their social media usage are valuable for estimating its causal impact on individuals, they overlook the broader impact social media may have on the collective behavior of one's entire network. In other words, social media may not only influence an individual because of the content they are exposed to online, but also because it changes the behavior of their social network. For example, social media could cause an individual's peer group to spend more time online and less time socializing in-person, which could have negative consequences for well-being.

Several states around the country are passing laws banning smartphones in schools, which provides us with a unique opportunity to measure the broader network-level effects of smartphone and social media usage. I have advocated for rigorous randomized control trials (RCTs) to measure the effectiveness of these bans (Rathje, *New York Times*, 2024), and am currently collaborating with partners to initiate an RCT on this topic. This RCT will help answer an important question: What are the consequences of an entire social network (as opposed to just an individual) spending time socializing in the online—rather than the offline—world?

Improving Methods for Global Research Using Artificial Intelligence

To improve methods for studying the psychology of virality, my work has also leveraged recent advances in artificial intelligence (e.g., the development of large-language models such as GPT) to better facilitate multi-lingual psychological text analysis. Recently, I demonstrated that GPT can accurately detect psychological constructs across 12 languages, including lesser-spoken languages, and that GPT is much more accurate than some other text analysis methods commonly used in the social sciences (Rathje et al., *PNAS*, 2024). We make the case that GPT will enable more global research that includes non-WEIRD (Western, Educated, Industrialized, Rich, and Democratic) and Global South populations, since it works well across several languages. It also democratizes advanced natural language processing techniques, making them more accessible to people around the globe.

A Global Account of the Psychology of Virality

I am using these recent methodological advances in multi-lingual psychological text analysis to come up with a more integrated theoretical account of what goes "viral" in online and offline social networks around the world. There are several competing theories about why content goes viral online (Rathje, Robertson, Brady, & Van Bavel, *Perspectives on Psychological Science*, 2024), and this study will put these theories to the test by examining the factors that best predict virality. I am currently analyzing various social media datasets (from Facebook, Instagram, LinkedIn, etc.) from dozens of countries around the globe using large language models. I am also testing for potential cross-cultural differences in what goes "viral" across different social media platforms and countries, building off prior work suggesting that information spreads differently in different cultural contexts (Hsu et al., 2021). Furthermore, I am testing whether the spread of hostile content online predicts important offline outcomes (e.g., violent conflict), globally.

In addition to exploring virality online, I plan to analyze *offline* conversations from around the world using these same methods to explore whether information spreads differently offline. Alongside this big data analysis, I am conducting controlled experiments to test whether the type of information people share online is different from the information they share offline. For instance, while out-group animosity goes viral online, is this also true in the context of gossip and word-of-mouth marketing? Or do certain affordances of the online world—such as algorithms, anonymity, and the "attention economy"—amplify the spread of hostile rhetoric? The goal of this research program is to develop a more comprehensive theoretical account of the psychology of virality in our online and offline networks.

Conclusion

Each day, millions of social media posts compete for our attention, yet only a few reach the top of our feeds. Some rumors die out, and others spread widely through whispered gossip. Understanding the psychology behind why information spreads is critical, especially as new technologies, such as social media and generative AI, rapidly change how information is consumed and shared. Understanding the consequences of exposure to (mis)information in our social media feeds and in our offline networks is equally crucial, particularly amid speculation that the rapid spread of "fake news" and AI "deepfakes" may be swaying elections, fostering intergroup conflict, and harming well-being. As a professor, I plan to integrate psychological theory with new methods—such as AI-based text analysis, global studies, and field experiments—to help us better understand these pressing questions.

References

Allport, G. W., & Postman, L. (1947). The psychology of rumor.

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational identity: A reader*, *56*(65), 9780203505984-16.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

Rathje, S., Van Bavel, J.J. & van der Linden, S. (2021) Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences.* https://doi.org/10.1073/pnas.2024292118

Heltzel, G., & Laurin, K. (2024). Why Twitter sometimes rewards what most people disapprove of: The case of cross-party political relations. *Psychological Science*.

Yu, X., Wojcieszak, M., & Casas, A. (2023). Partisanship on social media: In-party love among American politicians, greater engagement with out-party hate among ordinary users. *Political Behavior*, 1-26.

Rathje, S., Robertson, C., Brady, W. J., & Van Bavel, J. J. (2023). People think that social media platforms do (but should not) amplify divisive content. *Perspectives on Psychological Science*, 17456916231190392.

Rathje, S. & van der Linden, S. (2023). Shifting online incentive structures to reduce polarization and the spread of misinformation. *Research Handbook on Nudges and Society*.

Van Bavel, J.J., Robertson, C. del Rosario, K., Rasmussen, J., **Rathje, S.** Social Media and Morality (2023). *Annual Review of Psychology*.

Rathje, S., Roozenbeek, J., Van Bavel, J.J. & van der Linden, S. Accuracy and social motivations shape belief in (mis)information (2023). *Nature Human Behavior.* https://doi.org/10.1038/s41562-023-01540-w

Rathje, S., He, J., Roozenbeek, J., Van Bavel, J.J., & van der Linden, S. (2022). Social media behavior is associated with vaccine hesitancy. *Proceedings of the National Academy of Sciences – Nexus*. https://doi.org/10.1093/pnasnexus/pgac207

Rathje, S., He, J., Harjani, T., Roozebeek, J., Pretus, C., Gray, K., van der Linden, S, & Van Bavel, J.J. (In prep). The causal effect of social media (un)following behavior on affective polarization: results from a digital field experiment.

Van Bavel, J. J., **Rathje, S.**, Harris, E., Robertson, C., & Sternisko, A. (2021). How social media shapes polarization. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2021.07.013

Harris, E.*, **Rathje, S.***, Robertson, C., Van Bavel, J.J. (2023). The SPIR Model of Social Media and Polarization: Exploring the Role of Selection, Platform Design, Incentives, and Real-World Context. *International Journal of Communications*.
*co-first author

Rathje, S., Nejla, A., Robertson, C., Tucker, J., … Van Bavel, J.J. (In prep). The causal effect of social media usage around the globe. Invited as a Registered Report submission at *Nature*.

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature human behaviour*, *7*(1), 74-101.

Ghai, S., Fassi, L., Awadh, F., & Orben, A. (2023). Lack of sample diversity in research on adolescent depression and social media use: A scoping review and meta-analysis. *Clinical Psychological Science*, *11*(5), 759-772.

Rathje, S., Mirea*, D. M., Sucholutsky, I., Marjieh, R., Robertson, C., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Science*. doi/10.1073/pnas.2308950121
*co-first author

Rathje, S., To The Editor: Re "Why Schools Are Pushing Laws to Ban Smartphones" (2024). *The New York Times*. [*https://www.nytimes.com/2024/08/14/opinion/new-york-times-endorsements.html*](https://www.nytimes.com/2024/08/14/opinion/new-york-times-endorsements.html)

Hsu, T. W., Niiya, Y., Thelwall, M., Ko, M., Knutson, B., & Tsai, J. L. (2021). Social media users produce more affect that supports cultural values, but are more influenced by affect that violates cultural values. *Journal of Personality and Social Psychology*, *121*(5), 969.