Journal of Online Trust and Safety

Special Issue on: Uncommon yet Consequential Online Harms

Ronald E. Robertson

Introduction

A consistent finding from recent research on harmful online behaviors is that they tend to be concentrated among a small number of individuals. Whether examining misinformation (Grinberg et al. 2019; Guess, Nagler, and Tucker 2019; Allen et al. 2020), radicalizing or partisan content (Hosseinmardi et al. 2021; Muise et al. 2022), or hate speech (Zannettou et al. 2020), a small number of individuals typically accounts for the vast majority of the behavior. Despite this statistical infrequency, the social consequences of such behaviors can be substantial, especially for people targeted by, or within the social networks of, the individuals perpetuating them.

The observation that human behavior often follows a heavy-tailed distribution is well established (e.g. the Pareto principle), and applies to a wide range of behaviors, including the number of webpages people browse, papers scientists publish, and emails people receive (Newman 2005). In taking note of these patterns, scientists have studied their dynamics, proposed mechanisms for their origins, and identified structural features—such as "rich clubs," where the subgraphs of prominent individuals tend to be densely connected—to help us understand and reason about them (Colizza et al. 2006). However, in the context of harmful online behaviors, this facet of human behavior has rarely arisen as a focal point. While studies showing the low prevalence of such behaviors have been useful in countering alarmist narratives, their framing and focus on an "average user" can minimize the fact that small percentages can still constitute large absolute numbers, and that even a small number of individuals have the potential for an outsized impact (Guess 2021). Such concerns are further exacerbated by the connectivity of the Internet, which may enable such individuals to congregate on an unprecedented scale (Lorenz-Spreen et al. 2020), perhaps forming a different type of rich club.

In the push to summarize human behavior at scale, scholars have noted that "we risk losing sight of a secondary but equally important advantage of Big Data – the plentiful representation of minorities" (Foucault-Welles 2014). Similarly, behavioral researchers have noted that "a narrow focus on main effects in the population as a whole almost necessarily means a focus on effects in the group with the greatest numerical representation" (Bryan, Tipton, and Yeager 2021). While the case for "making big data small" (Foucault-Welles 2014) has often been focused on groups historically omitted from the scientific record, it may also lend itself to studying or intervening on the individuals perpetuating the patterns outlined here. So, the question is: we know this pattern of uncommon yet consequential online harms exists, now what?

doi:10.54501/jots.v1i3.87

Issue Outline

In this special issue we present four papers that examine various aspects of our theme: uncommon yet consequential online harms, a phrase inspired by a related discussion of "events that are statistically uncommon but consequential" in Lazer et al. (2021).

In "Election Fraud, YouTube, and Public Perception of the Legitimacy of President Biden," authors James Bisbee, Megan A. Brown, Angela Lai, Richard Bonneau, Joshua A. Tucker, and Jonathan Nagler shed light on the role of YouTube's recommender system in driving skeptical users to videos that featured fraud-related narratives about the 2020 US Elections. Using an algorithm auditing method, Bisbee et al recruited participants to complete an opinion survey, install a browser extension, and follow YouTube recommendations based on a randomly chosen traversal rule (e.g. always click on the second recommendation) from a randomly chosen seed video. They find that recommendations to videos featuring election-fraud related narratives were uncommon overall, and largely concentrated among a small number of individuals who self-reported high skepticism about the election's legitimacy (Bisbee et al. 2022).

In "Predictors of Radical Intentions among Incels: A Survey of 54 Self-identified Incels," authors Sophia Moskalenko, Naama Kates, Juncal Fernández-Garayzábal González, and Mia Bloom conducted a survey of self-identified "incels" ("involuntary celibates") that focused on mental health, incel ideology, and radical intentions. Moskalenko et al. studied this small online-based identity group, of which a small number have committed heinous real-world crimes, through an innovative recruitment approach that involved surveying incels who sought to speak with an interventionist at a nonprofit in the US. Their results show high rates of self-reported mental health issues among their self-selected participants, like depression and anxiety, but low rates of radical attitudes or intentions (Moskalenko et al. 2022).

In "Procedural Justice and Self Governance on Twitter: Unpacking the Experience of Rule Breaking on Twitter," authors Matthew Katsaros, Tom Tyler, Jisu Kim, and Tracey Meares investigated how Twitter users who violated the platform's content rules perceived and responded to the enforcement action (i.e. post removal). In collaboration with Twitter, Katsaros et al. used surveys paired to behavioral data to examine whether those who received the enforcement action subsequently changed their subsequent behavior on the platform, and how their subsequent behavior related to their self-reported perceptions of procedural justice. Their research sheds light on the motivations and perceptions of those who break Twitter's rules, shows that those who felt they were treated more fairly were less likely to recidivate, and shows that only a small minority report breaking the rules with the specific aim of harming someone (Katsaros et al. 2022).

In "Twitter's Disputed Tags May Be Ineffective at Reducing Belief in Fake News and Only Reduce Intentions to Share Fake News Among Democrats and Independents," authors Jeffrey Lees, Abigail McCarter, and Dawn M. Sarno conducted a survey-based experiment to measure the potential impact of the "This claim is disputed" tags placed on posts deemed to be misinformation. More specifically, Lees et al. examine how their participants' reactions to these tags vary based on their self-reported demographic and psychographic characteristics. Overall, their results show mixed evidence for the impact of such tags and suggest that they may have only reduced the likelihood of sharing false information for Democrats and Independents, but not Republicans, a finding consistent with prior work that suggests a small number of Republicans account for most fake news sharing within online platforms (Lees, McCarter, and Sarno 2022).

Conclusion

This special issue joins recent work by psychologists and behavioral scientists in asking the research community to reconsider how they think about effect sizes, interventions, and heterogeneity (Bryan, Tipton, and Yeager 2021; Funder and Ozer 2019; Lorenz-Spreen et al. 2020). Without such considerations, we may overestimate the impact of the interventions we design, and miss opportunities to evaluate how they might be improved in ways that account for heterogeneous responses (Szaszi et al. 2022) or cumulative effects (Abelson 1985). As noted in the “curb-cut effect”—which posits that “laws and programs designed to benefit vulnerable groups, such as the disabled or people of color, often end up benefiting all of society” (Blackwell 2016)—adjusting systems for a small number of individuals can often have positive effects for the entire population. Perhaps the same will prove true here. We hope that this special issue inspires future research in this area, and that such research paves the way for a deeper understanding and more effective solutions to uncommon yet consequential online harms.

References

Abelson, Robert P. 1985. "A Variance Explanation Paradox: When a Little Is a Lot." *Psychological Bulletin* (US) 97 (1): 129–33. https://doi.org/10.1037/0033-2909.97.1.129.

Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. 2020. "Evaluating the Fake News Problem at the Scale of the Information Ecosystem." *Science Advances* 6 (14). https://doi.org/10.1126/sciadv.aay3539.

Bisbee, James, Megan A. Brown, Richard Bonneau, Joshua A. Tucker, and Jonathan Nagler. 2022. "Election Fraud, YouTube, and Public Perception of the Legitimacy of President Biden." *Journal of Online Trust and Safety* 1 (3). https://doi.org/10.54501/jots.v1i3.60.

Blackwell, Angela Glover. 2016. "The Curb-Cut Effect." *Stanford Social Innovation Review* 15:2833. https://doi.org/10.48558/YVMS-CC96.

Bryan, Christopher J., Elizabeth Tipton, and David S. Yeager. 2021. "Behavioural Science Is Unlikely to Change the World without a Heterogeneity Revolution." *Nature Human Behaviour* 5, no. 8 (August): 980–89. https://doi.org/10.1038/s41562-021-01143-3.

Colizza, V., A. Flammini, M. A. Serrano, and A. Vespignani. 2006. "Detecting Rich-Club Ordering in Complex Networks." *Nature Physics* 2, no. 2 (February): 110–15. https://doi.org/10.1038/nphys209.

Foucault-Welles, Brooke. 2014. "On Minorities and Outliers: The Case for Making Big Data Small." *Big Data & Society* 1, no. 1 (April): 205395171454061. https://doi.org/10.1177/2053951714540613.

Funder, David C., and Daniel J. Ozer. 2019. "Evaluating Effect Size in Psychological Research: Sense and Nonsense." *Advances in Methods and Practices in Psychological Science* 2, no. 2 (June): 156–68. https://doi.org/10.1177/2515245919847202.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. "Fake News on Twitter during the 2016 U.S. Presidential Election." *Science* 363 (6425): 374–78. https://doi.org/10.1126/science.aau2706.

Guess, Andrew M. 2021. "(Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets." *American Journal of Political Science,* https://doi.org/10.1111/ajps.12589.

Guess, Andrew M., Jonathan Nagler, and Joshua Tucker. 2019. "Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook." *Science Advances* 5 (1). https://doi.org/10.1126/sciadv.aau4586.

Hosseinmardi, Homa, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M. Rothschild, and Duncan J. Watts. 2021. "Examining the Consumption of Radical Content on YouTube." *Proceedings of the National Academy of Sciences* 118 (32). https://doi.org/10.1073/pnas.2101967118.

Katsaros, Matthew, Tom Tyler, Jisu Kim, and Tracey Meares. 2022. "Procedural Justice and Self Governance on Twitter: Unpacking the Experience of Rule Breaking on Twitter." *Journal of Online Trust and Safety* 1 (3). https://doi.org/10.54501/jots.v1i3.38.

Lazer, David, Eszter Hargittai, Deen Freelon, Sandra Gonzalez-Bailon, Kevin Munger, Katherine Ognyanova, and Jason Radford. 2021. “Meaningful Measures of Human Society in the Twenty-First Century.” *Nature* 595, no. 7866 (July): 189–96. https://doi.org/10.1038/s41586-021-03660-7.

Lees, Jeffrey, Abigail McCarter, and Dawn M. Sarno. 2022. “Twitter’s Disputed Tags May Be Ineffective at Reducing Belief in Fake News and Only Reduce Intentions to Share Fake News Among Democrats and Independents.” *Journal of Online Trust and Safety* 1 (3). https://doi.org/10.54501/jots.v1i3.39.

Lorenz-Spreen, Philipp, Stephan Lewandowsky, Cass R. Sunstein, and Ralph Hertwig. 2020. “How Behavioural Sciences Can Promote Truth, Autonomy and Democratic Discourse Online.” *Nature Human Behaviour* 4, no. 11 (November): 1102–9. https://doi.org/10.1038/s41562-020-0889-7.

Moskalenko, Sophia, Naama Kates, Juncal Fernández-Garayzábal González, and Mia Bloom. 2022. “Predictors of Radical Intentions among Incels: A Survey of 54 Self-identified Incels.” *Journal of Online Trust and Safety* 1 (3). https://doi.org/10.54501/jots.v1i3.57.

Muise, Daniel, Homa Hosseinmardi, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. 2022. “Quantifying Partisan News Diets in Web and TV Audiences.” *Science Advances* 8, no. 28 (July): eabn0083. https://doi.org/10.1126/sciadv.abn0083.

Newman, M.E.J. 2005. “Power Laws, Pareto Distributions and Zipf’s Law.” *Contemporary Physics* 46, no. 5 (September): 323–51. https://doi.org/10.1080/00107510500052444.

Szaszi, Barnabas, Anthony Higney, Aaron Charlton, Andrew Gelman, Ignazio Ziano, Balazs Aczel, Daniel G. Goldstein, David S. Yeager, and Elizabeth Tipton. 2022. “No Reason to Expect Large and Consistent Effects of Nudge Interventions.” *Proceedings of the National Academy of Sciences* 119, no. 31 (August): e2200732119. https://doi.org/10.1073/pnas.2200732119.

Zannettou, Savvas, Mai Elsherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. “Measuring and Characterizing Hate Speech on News Websites.” In *12th ACM Conference on Web Science,* 125–34. Southampton United Kingdom: ACM, July. https://doi.org/10.1145/3394231.3397902.

Authors

Ronald E. Robertson is a Postdoctoral Fellow at the Stanford Internet Observatory (SIO) and Guest Editor at the *Journal of Online Trust and Safety* (ronalder@stanford.edu).

Acknowledgements

I'm grateful to Shelby Grossman, Dan Bateyko, and the rest of the SIO staff for their advice, help, and guidance in putting together this special issue and its accompanying symposium, which was held earlier this year. I also thank the reviewers for their valuable time and expertise in evaluating these papers, and the authors for their thoughtful contributions to this special issue.

Data Availability Statement

Not applicable.

Funding Statement

Not applicable.

Ethical Standards

Not applicable.

Keywords

Online behavior; misinformation; hate speech; radicalization.