The Digital Engine of Democracy: How Social Media Algorithms Drive Political Polarization

Word Count: 1141

A paper submitted to College Board for AP Seminar
April 9, 2026

Introduction

The manipulation of public opinion via digital platforms has become a critical threat to modern democracy. As political discourse has migrated online over the years, so have the key mechanisms of political persuasion. A 2018 global inventory by the Oxford Internet Institute found formally-organized social media manipulation campaigns in 48 countries, revealing a multi-million dollar industry dedicated to internet propaganda and disinformation (Bradshaw & Howard, 2018). While state-sponsored propaganda is well documented, an equal if not greater threat lies in the strategies used by domestic political campaigns and the inherent architectures of social media platforms themselves. This report investigates the intersection of algorithms, psychological microtargeting strategies, and digital campaign financing, exploring how the modern political landscape exploits user psychology to drive both ideological polarization and capital extraction.

The Psychology of Algorithmic Outrage and Microtargeting

To understand the efficacy of digital political manipulation, we must first examine how social media platforms leverage human psychology. In their peer-reviewed behavioral study, Brady et al. (2021) demonstrate that social media platforms amplify moral outrage through social reinforcement learning. The authors argued that because platforms reward moral and emotional expressions with highly quantifiable social feedback (likes, shares, and comments), users are continuously conditioned to increase their outrage expressions. Their methodology, encompassing the analysis of 12.7 million tweets alongside controlled behavioral experiments, provides evidence that social media algorithms do not serve as neutral channels; rather, they actively shape moral discourse by exploiting natural human tendencies of social conformity.

Political campaigns purposefully weaponize these psychological vulnerabilities through Political Microtargeting (known as PMT). Zarouali et al. (2022) conducted an experimental study demonstrating that political advertisements are significantly more persuasive when tailored to a user's personality traits, such as extroversion or introversion. By utilizing a mock social network-

ing site and automated personality profiling algorithms, the authors establish that when citizens are targeted with emotionally charged ads matching their psychological profiles (e.g. enthusiasm for extroverts, fear for introverts), their voting intentions increase. This reveals a dangerous asymmetry wherein campaigns can exploit voters' deeply rooted psychological dispositions for political gain (Zarouali et al., 2022).

The Efficacy and Limitations of Platform Interventions

Given the clear psychological exploitation facilitated by social media feeds, a logical counter-perspective suggests that reforming platform algorithms could cure digital polarization. However, a pair of landmark 2023 studies published in Science complicate this view. Guess et al. (2023a) conducted a massive randomized controlled experiment during the 2020 U.S. election, replacing Facebook and Instagram's machine-learning algorithms with a simple reverse-chronological feed for a subset of users. Concurrently, Guess et al. (2023b) tested the removal of the "reshare" function.

As peer-reviewed studies authored by a coalition of independent academics with unprecedented platform access, these papers carry immense credibility. While removing algorithms and reshares drastically reduced the time users spent on the platforms and increased their exposure to untrustworthy content (the chronological feed paradoxically surfaced more low-quality partisan spam than the algorithm), the researchers found no statistically significant changes in users' thought polarization, nor in their political knowledge (Guess et al., 2023a; Guess et al., 2023b).

Taking these perspectives into account, a nuanced reality emerges: while algorithms undoubtedly reward and amplify outrage (as established by Brady et al.), they are not the sole source of political polarization. This is further corroborated by recent experimental evidence regarding video recommendation systems. Liu et al. (2025) experimentally manipulated YouTube's algorithm to simulate extreme "filter bubbles" and "rabbit holes" for nearly 9,000 users, finding no detectable polarizing effects on short-term political attitudes. As Guess et al. and Liu et al. collectively sug-

gest, simply reverting to chronological feeds or re-engineering recommendations does not reverse deep-seated societal divides, indicating that while platforms are definitely catalysts, humans are inherently predisposed to polarization. Furthermore, expecting platforms to effectively self-regulate is mostly unrealistic. An internal investigation by The Wall Street Journal revealed that Facebook secretly operated a program called "XCheck," which exempted millions of high-profile users (including politicians) from standard moderation rules, allowing them to post inflammatory or false content without consequence (Horwitz, 2021). This exposes a fundamental platform hypocrisy, undermining the premise that corporate social media entities can serve as reliable and neutral arbiters of democratic discourse.

The Financial Weaponization of the Digital Medium

The ultimate manifestation of algorithmic outrage and targeted polarization is financial. Emotional stimulation, particularly anger, has become a primary driver of modern campaign finance. Chapman et al. (2022) investigate the phenomenon of "rage donations," hypothesizing that exposure to opposing views provokes anger and subsequent collective giving. While their lab experiments found mixed results regarding opposition-induced rage, they established that supportive advocacy mobilizes giving by enhancing a sense of efficacy and group identification.

This mobilization is highly facilitated by frictionless campaign technologies like ActBlue and WinRed, which seamlessly convert ideological favor into capital. Yorgason (2024) argues that the adoption of ActBlue by Democratic candidates directly causes a shift toward ideological extremism. Because ActBlue drastically lowers the barrier to entry for un-itemized, small-dollar donors, (a demographic historically proven to be more ideologically extreme than the general electorate), candidates become financially beholden to this base. Consequently, Yorgason shows that lawmakers who heavily utilize ActBlue shift their congressional voting records further to the left to appease these polarized donors. The real-world efficacy of this strategy is evident in modern progressive fundraising. For example, Representative Alexandria Ocasio-Cortez and New York

Mayor Zohran Mamdani have utilized small-dollar, ideologically driven networks to out-fundraise establishment politicians by massive margins (Kanno-Youngs, 2025).

However, the pursuit of digital campaign capital frequently crosses ethical boundaries. Posner et al. (2023) detail the deployment of "dark defaults" by Republican campaigns using the WinRed platform during the 2020 election. By hiding pre-checked boxes in their user interfaces, campaigns automatically opted donors into weekly recurring donations. Posner et al.'s analysis of Federal Election Commission refund data provides quantitative evidence of this deception, revealing that these dark defaults generated over $43 million in unintended donations, subsequently leading to massive spikes in refund requests from defrauded citizens. This highlights how architecture is weaponized to exploit the same behavioral vulnerabilities utilized by most social media algorithms.

Conclusion

The architecture of the modern digital "town square" represents a cooperative exploitation of human behavior. As Bradshaw and Howard (2018) established, the digital manipulation of voters is a global, institutionalized practice. While algorithmic amplification (Brady et al., 2021) and personality microtargeting (Zarouali et al., 2022) actively cultivates and directs voter outrage, eliminating these algorithms does not instantly cure the electorate's deep-seated polarization (Guess et al., 2023a). Instead, this psychological priming is seamlessly monetized by digital campaign infrastructures. By reducing friction (Yorgason, 2024) and employing deceptive UI tactics (Posner et al., 2023), campaigns convert manufactured outrage into unprecedented financial war chests, ultimately incentivizing politicians to adopt increasingly extreme legislative stances. Therefore, mitigating digital polarization requires looking beyond simple algorithmic tweaks, demanding a rigorous regulation of the mechanisms that currently define online political engagement.

References

Bradshaw, S., & Howard, P. N. (2018). Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation. Computational Propaganda Research Project, Oxford Internet Institute.

Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. Science Advances, 7(33). https://doi.org/10.1126/sciadv.abe5641

Chapman, C. M., Lizzio-Wilson, M., Mirnajafi, Z., Masser, B. M., & Louis, W. R. (2022). Rage donations and mobilization: Understanding the effects of advocacy on collective giving responses. British Journal of Social Psychology, 61(3), 882–906. https://doi.org/10.1111/bjso.12522

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Velasco Rivera, C., Settle, J., Thomas, D. R., … Tucker, J. A. (2023a). How do social media feed algorithms affect attitudes and behavior in an election campaign? Science, 381(6656), 398–404. https://doi.org/10.1126/science.abp9364

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Velasco Rivera, C., Settle, J., Thomas, D. R., … Tucker, J. A. (2023b). Reshares on social media amplify political news but do not detectably affect beliefs or opinions. Science, 381(6656), 404–408. https://doi.org/10.1126/science.add8424

Horwitz, J. (2021, September 13). Facebook says its rules apply to all. Company documents reveal a secret elite that's exempt. The Wall Street Journal. https://www.wsj.com/tech/facebook-files-xcheck-zuckerberg-elite-rules-11631541353

Kanno-Youngs, Z. (2025, July 15). AOC, Mamdani and progressives rake in cash as Democrats remain divided. The Wall Street Journal. https://www.wsj.com/politics/elections/democratic-donors-progressive-aoc-zohran-mamdani-825242e9

Liu, N., Hu, X. E., Savas, Y., Baum, M. A., Berinsky, A. J., Chaney, A. J. B., Lucas, C., Mariman, R., de Benedictis-Kessner, J., Guess, A. M., Knox, D., & Stewart, B. M. (2025). Short-term exposure to filter-bubble recommendation systems has limited polarization effects: Naturalistic experiments on YouTube. Proceedings of the National Academy of Sciences, 122(8), e2318127122. https://doi.org/10.1073/pnas.2318127122

Posner, N., Simonov, A., Mrkva, K., & Johnson, E. J. (2023). Dark defaults: How choice architecture steers political campaign donations. Proceedings of the National Academy of Sciences, 120(40). https://doi.org/10.1073/pnas.2218385120

Yorgason, C. (2024). Darker Blue: How Small Donors Drive Congressional Polarization. Stanford University.

Zarouali, B., Dobber, T., De Pauw, G., & de Vreese, C. (2022). Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media. Communication Research, 49(8), 1066–1091. https://doi.org/10.1177/0093650220961965