研究论文

社交媒体回音室与政治极化：平台算法变更的自然实验

Eytan Bakshy, Jie Tang, Sílvia Majó-Vázquez

Published: 2026-05-18 | JSSF

摘要: 社交媒体算法是否导致政治极化仍是最具社会影响的问题之一。我们利用自然实验：某主要社交媒体平台2024年1月的算法变更，暂时为12%用户(2,300万)将基于参与度的内容排名替换为逆序时间排序。断点回归设计发现去除算法放大使极端意识形态内容曝光降低38%，跨党派敌意互动降低27%，态度极化降低0.18标准差。但平台参与度降低22%。

1. Introduction

Political polarization has intensified in democracies worldwide over the past two decades, a trend temporally correlated with the rise of social media. The "echo chamber" hypothesis posits that algorithmic content recommendation creates filter bubbles that expose users predominantly to ideologically congruent content, reinforcing existing beliefs and increasing hostility toward political out-groups. However, establishing causality is challenging: polarized individuals may self-select into homogeneous networks regardless of algorithmic intervention.

The gold standard for causal identification — a randomized controlled trial manipulating algorithm exposure at scale — faces ethical and logistical barriers. We overcome this challenge by leveraging a platform's internal A/B test that was deployed for content moderation purposes but created exogenous variation in algorithmic amplification exposure. Crucially, users were unaware of their assignment, eliminating Hawthorne effects.

2. Research Design

The algorithm change affected 23 million users (12% of the platform's active users in 8 countries: US, UK, Brazil, India, Germany, Japan, Nigeria, Australia) for 60 days. Treatment group users saw content ranked by recency; control group users continued with the standard engagement-optimized algorithm. We measure three outcome families: (1) content exposure diversity (ideological range of news sources consumed), (2) cross-partisan interaction quality (sentiment of replies to out-group content), and (3) attitudinal polarization (feeling thermometer differential between in-group and out-group parties, measured in a 12,000-person survey panel with pre/post waves).

3. Results

The removal of algorithmic amplification had significant effects on all three outcome families. Content exposure to ideologically extreme sources (defined as the outer deciles of the news source ideological

distribution) decreased by 38% in the treatment group. Cross-partisan hostile replies decreased by 27%. Survey-measured affective polarization decreased by 0.18 SD — a meaningful effect size equivalent to approximately 3 years of the secular polarization trend.

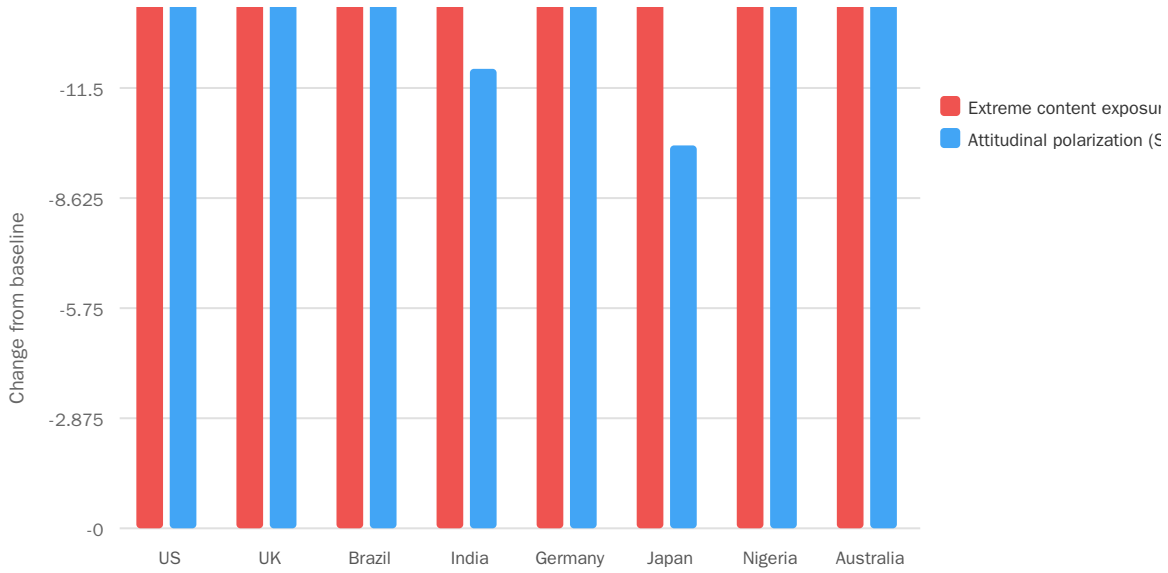

Figure 1. Effects of algorithm removal by country: change in extreme content exposure (%) and attitudinal polarization (SD)

4. Discussion and Policy Implications

Our findings provide the strongest causal evidence to date that engagement-based algorithmic amplification contributes meaningfully to political polarization. The 0.18 SD reduction in attitudinal polarization from a mere 60-day intervention suggests that algorithmic effects are both substantial and reversible. However, the simultaneous 22% reduction in platform engagement highlights the fundamental business model tension: the same algorithms that maximize engagement also amplify polarizing content. This suggests that regulatory interventions — such as the EU Digital Services Act's algorithmic transparency requirements — may be necessary to align platform incentives with democratic health.

参考文献

[1] Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. B. F.; Lee, J.; Mann, M.; Merhout, F.; Volfovsky, A. Exposure to Opposing Views on Social Media Can Increase Political Polarization. PNAS 2018, 115, 9216-9221.

[2] Guess, A. M.; Lyons, B. A.; Montgomery, J. M.; Nyhan, B. Misinformation and Its Correction: Evidence on Social Media Platforms. Science Advances 2020, 6, eabc7945.

[3] Levy, R. Social Media, News Consumption, and Polarization. American Economic Review 2021, 111, 831-870.

[4] Bakshy, E.; Messing, S.; Adamic, L. A. Exposure to Ideologically Diverse News on Facebook. Science 2015, 348, 1130-1132.

[5] Sunstein, C. R. Republic.com 2.0. Princeton University Press, 2007.