Engagement-based algorithms disrupt social norm learning

Meriel Doyle[1], Silvan Baier[1], Joshua C. Jackson[2] & William J. Brady[1]

[1]Kellogg School of Management, Northwestern University
[2]Booth School of Business, University of Chicago

(Q1) Which categories of content are amplified by engagement-based algorithms?

(Q2) Do engagement-based algorithms disrupt social norm learning?

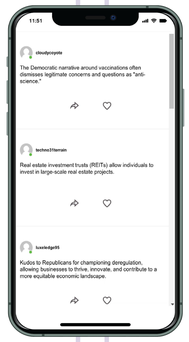
(Q3) Can bridging-based algorithms correct social norm learning?

Methods

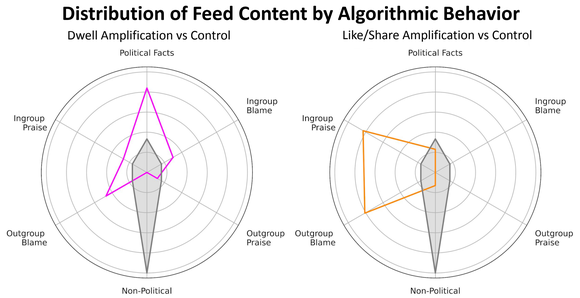
Study: Three online experimental studies (total *N* = 3500) where participants scrolled through a mock social media newsfeed.

Study features:

- Feeds consisted of validated GPT-generated tweets that varied on ingroup, moral, emotional dimensions ('PRIME information'; Brady et al., 2023)
- Measured **dwell time** (how much time people spend viewing each message on their screen) and **likes/shares** of messages
- Generated new feeds based on: (1) **Dwell** (upranked posts w/ highest dwell time), (2) **Like/Share** (upranked posts w/ highest like/share frequency), (3) **Bridging** (upranked posts liked by *both* democrats and republicans), (4) **Control** (random sample of original content distribution)



Key Findings (Q1)



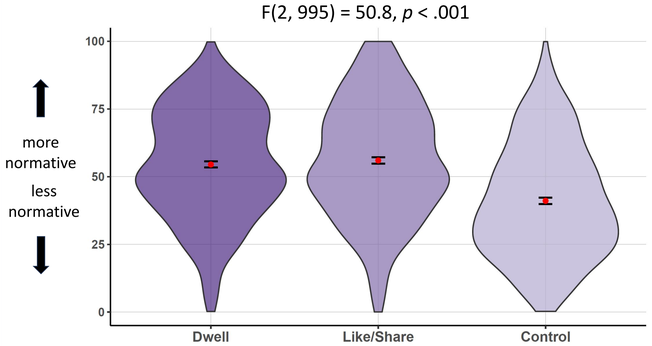
Outgroup blame, ingroup praise, political content (*ingroup slanted content*) are amplified by engagement-based algorithms (Study 1)

Key Findings (Q2)

Perceived Commonality of Blame

"What percentage (%) of people in the social network you viewed were posting a message that blamed others?"

F(2, 995) = 50.8, $p < .001$



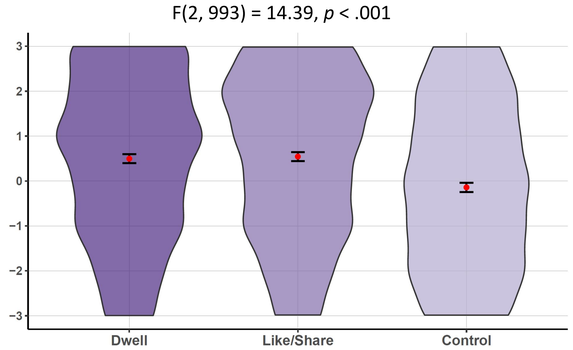
Dwell vs. Control: MD = -13.43, $p < .001$, $d = 0.65$

Like/Share vs. Control: MD = -14.92, $p < .001$, $d = 0.70$

Perceived Normativity of Blame

"How socially appropriate do you think it is to post a message that blames others in the social network that you just viewed?"

F(2, 993) = 14.39, $p < .001$



Dwell vs. Control: MD = -0.64, $p < .001$, $d = 0.35$

Like/Share vs. Control: MD = -0.68, $p < .001$, $d = 0.37$

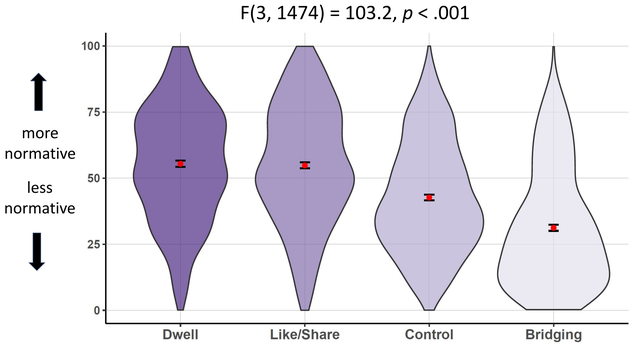
Engagement-based algorithms *overrepresent* social normativity of blame in online social networks (Study 2)

Key Findings (Q3)

Perceived Commonality of Blame

"What percentage (%) of people in the social network you viewed were posting a message that blamed others?"

F(3, 1474) = 103.2, $p < .001$



Dwell vs. Control: MD = -12.79, $p < .001$, $d = 0.60$

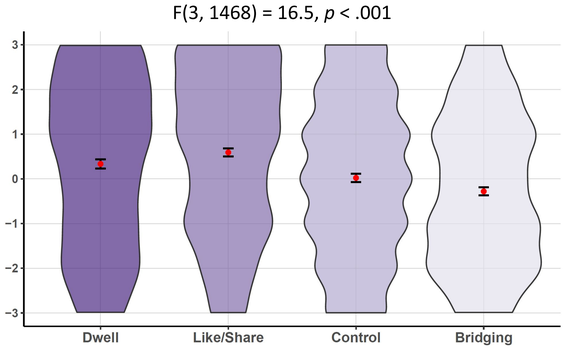
Like/Share vs. Control: MD = -12.08, $p < .001$, $d = 0.56$

Bridging vs. Control: MD = -11.44, $p < 0.001$, $d = 0.53$

Perceived Normativity of Blame

"How socially appropriate do you think it is to post a message that blames others in the social network that you just viewed?"

F(3, 1468) = 16.5, $p < .001$



Dwell vs. Control: MD = -.32, $p = .023$, $d = 0.17$

Like/Share vs. Control: MD = -0.57, $p < .001$, $d = 0.31$

Bridging vs. Control: MD = -0.31, $p = .023$, $d = 0.18$

Bridging-based algorithms decrease social normativity of blame but *underrepresent* true distribution of content (Study 3)