dreammachine

SYSTEMS ARCHITECTURE FOR THE HUMAN MIND

HALLUCINATION AUDIT

EMPIRICAL TRACEABILITY LOG

Permalink · dreammachine.systems/audit/sec7-1-1

LOG ID	BOOK REFERENCE	RCT FILTER	CONFIDENCE TIER
SEC 7.1.1	CHAPTER 7: STANDARDIZING CONNECTION (Social API)	**OFF**	**CONSENSUS**

AUDITED CLAIM

PREDICTIVE EXPLOITATION (SOCIAL DDOS ATTACK)

SYSTEMS MODEL — AS STATED IN THE BOOK

> Empirically, social media predictive ranking systems **maximize user engagement** by disproportionately amplifying content that triggers primal, high-arousal negative emotions like **moral outrage** and **out-group hostility**.

VALIDATION QUERY

Do social media predictive algorithms maximize user engagement by amplifying content that triggers fear and moral outrage?

EMPIRICAL FINDING

Independent algorithmic audits and computational social science establish that predictive ranking algorithms create a feedback loop that **maximizes user engagement** by disproportionately amplifying content evoking **fear**, **moral outrage**, and **out-group hostility** because these primal triggers most reliably capture human attention.

PRIMARY ANCHORS

→ Rathje et al., 2021, Proceedings of the National Academy of Sciences

→ Guess et al., 2023, Science

ABOUT THIS RECORD

Each claim in the book is decompiled from its IT metaphor into a naked, falsifiable scientific claim, then anchored to a **primary peer-reviewed source**. Claim and anchor are queried against the Consensus academic database (200M+ papers) under a fixed protocol — top-tier **Q1/Q2 journals only**, preprints excluded, human studies only. If a hypothesis fails the audit, the build is rejected: the AI generates the syntax; the peer-reviewed grid validates the physics.

CONFIDENCE TIERS

- **CONSENSUS ◀ this log** — broadly established, highly replicated mechanics.
- **EMERGING** — plausible, actively researched; no universal consensus yet.
- **EXPERIMENTAL** — contested, extrapolated, or purely theoretical.

RCT FILTER — Controlled Studies is ON for interventions (a forced action must beat a control) and OFF for mechanisms (so foundational neuroscience isn't blocked).

Yes, **social media algorithms amplify fear and moral outrage** because that content tends to **drive engagement**.

1. Introduction

Social media predictive algorithms appear to maximize user engagement partly by amplifying content that triggers fear, moral outrage, and out-group hostility. The strongest evidence does not show a single direct proof that platforms explicitly optimize for “fear and outrage” as a formal objective, but it does show a consistent chain: emotionally and morally charged content captures attention, gets more engagement, algorithms reward engagement, and changing ranking systems changes exposure to uncivil or partisan-animosity content (Guess et al., 2023; Piccardi et al., 2024; McLoughlin & Brady, 2023).

Large observational studies show that moral-emotional language increases diffusion, that out-group animosity is one of the strongest predictors of sharing, and that outrage expression is reinforced by social feedback in platform environments (Brady et al., 2017; Rathje et al., 2021; Brady et al., 2021). Reviews and theory papers converge on the same mechanism: platforms are built to capture attention, and moralized, threatening, or divisive content is especially good at doing so (Van Bavel et al., 2023; McLoughlin & Brady, 2023).

The important caveat is that evidence for downstream attitude change is weaker than evidence for engagement effects. Several causal studies find that algorithmic ranking affects what people see and how much they engage, but short-term changes in ranking do not always produce large changes in polarization or policy attitudes (Guess et al., 2023; Liu et al., 2025).

Do social media predictive algorithms maximize user engagement by amplifying content that triggers fear and moral outrage?
Requires at least 5 papers that directly answer your question. Try adjusting your query to find more papers.

FIGURE 1 Research consensus on algorithmic amplification of outrage

2. Methods

The search ran over more than 170 million research papers indexed by Consensus, drawing from Semantic Scholar, PubMed, and other scholarly sources. It identified 5,739 directly relevant papers in the initial query stream, expanded through targeted searches on emotional amplification, recommender systems, misinformation, polarization, fear appeals, and moral outrage, then screened 240 papers and selected 50 for this synthesis.

Search Strategy

21.1M
Retrieved

→

1.9K
Eligible

→

50
Included

22 searches

1 citation graph use

All papers retrieved from 22 searches and 1 citation graph use.

Papers meeting relevance and quality criteria, after deduplication

Top papers selected to be included in the final analysis

FIGURE 2 Deep search screening and inclusion flow

The strategy combined foundational theory, mechanism-focused queries, critique searches, and adjacent-topic searches to capture both supportive and contradictory evidence.

3. Results

3.1 Key Papers

A few papers anchor the literature because they establish the main empirical links between emotional content, engagement, and algorithmic curation. The foundational work shows that moral-emotional language spreads more widely, out-group hostility strongly predicts engagement, and feed-ranking interventions causally change exposure and engagement patterns (Brady et al., 2017; Rathje et al., 2021; Guess et al., 2023).

Paper	Summary
(Piccardi et al., 2024)	Algorithmic feeds increase time spent and alter exposure patterns (Guess et al., 2023)
(Brady et al., 2019)	Out-group animosity predicts shares better than negative affect (Rathje et al., 2021)
(Chang et al., 2025)	Moral-emotional words increase diffusion by 20% each (Brady et al., 2017)

FIGURE 3 Foundational papers on engagement and emotional diffusion

3.2 Engagement Mechanisms

Emotionally charged content spreads faster and farther than neutral content in multiple social media settings (Stieglitz & Dang-Xuan, 2013; Brady et al., 2019). Moral-emotional words increase retweets by about 20% per word in political discussion, and similar effects appear among political elites (Brady et al., 2017; Brady et al., 2019). Out-group terms raise the odds of sharing by 67%, and posts about political out-groups are shared about twice as often as in-group posts (Rathje et al., 2021).

Fear works through a similar attention pathway. Moderate fear-arousing sensationalism increased Facebook engagement during the Zika outbreak, and virality cues made identical threat-related content feel more dangerous and more worth responding to with outrage (Ali et al., 2019; Puryear et al., 2024).

3.3 Algorithmic And Social Reinforcement

Feature	Result
Study type	Preregistered observational and experimental studies (Brady et al., 2021)
Sample	7,331 users, 12.7 million tweets, plus experiments (Brady et al., 2021)
Core mechanism	Social feedback predicts future outrage expression (Brady et al., 2021)
Network effect	Norms in users' networks shape outrage expression (Brady et al., 2021)
Algorithm link	Feeds affect feedback by controlling exposure (Brady et al., 2021)
Implication	Engagement-oriented design can indirectly amplify outrage (Brady et al., 2021)

FIGURE 4 Landmark study on outrage reinforcement mechanisms

The key takeaway is that algorithms need not "prefer outrage" explicitly to increase it; amplifying feedback-rich content is enough (Brady et al., 2021).

3.4 Causal Tests And Limits

The best causal evidence shows ranking algorithms change exposure and engagement, but effects on beliefs are more variable. In Facebook and Instagram field experiments during the 2020 US election, switching to chronological feeds reduced time spent and activity, decreased uncivil content on Facebook, but did not significantly change major political attitudes over three months (Guess et al., 2023). In a separate field experiment on X, reranking partisan animosity content changed affective polarization, and reducing such content appeared to reduce short-term engagement, which supports the idea that highly provocative content is engagement-efficient (Piccardi et al., 2024).

At the same time, short-term YouTube experiments found limited effects of manipulated recommendation slant on policy attitudes, and a review of filter-bubble claims concluded the broader empirical basis for strong algorithmic-polarization claims remains uneven (Liu et al., 2025; Bruns, 2019).

Results Timeline

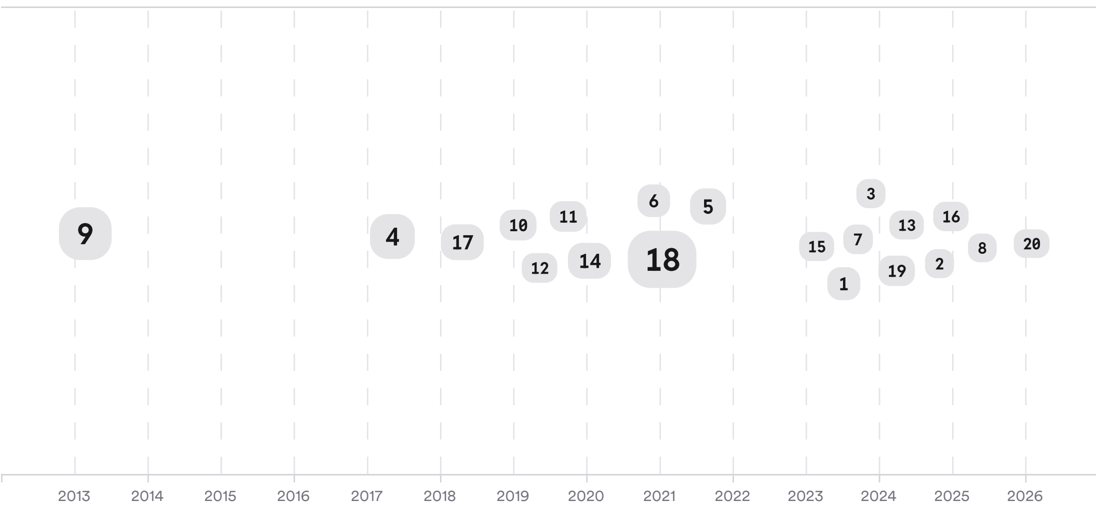

FIGURE 5 Timeline of influential findings with citation-weighted markers

Early studies established that emotional content diffuses more efficiently online, mid-period work identified moral contagion and out-group animosity as especially powerful, and recent field experiments began testing ranking interventions directly (Stieglitz & Dang-Xuan, 2013; Brady et al., 2017; Rathje et al., 2021; Piccardi et al., 2024). The trajectory is moving from correlational virality studies toward causal auditing of platform ranking systems (Piccardi et al., 2024; Liu et al., 2025).

Top Contributors

Type	Name	Papers
Author	W. Brady	[881a0ce0edc4573d82bb4635a5487fb8] [b323dc32aee0542b91c73b6fe8db6bf0] [7efe43bf09d15b0a9e358dfbf394ae5c] [b7ed8d675a2e5ab8b0d15778587f3c23]
Author	Jay J. Van Bavel	[d064335f1cb45d0aa03c3040663b1ae5] [4366815bd48757cd96598e4a18256b69] [238bf96c63a750f680d5274a1815595e] [7efe43bf09d15b0a9e358dfbf394ae5c]
Author	Killian L. McLoughlin	[881a0ce0edc4573d82bb4635a5487fb8] [37d6d3fdace05eeda06b3b5e1498e768] [3d6b81802a7d57b69aa9b2e06aae314f] [b7ed8d675a2e5ab8b0d15778587f3c23]

Type	Name	Papers
Journal	*Proceedings of the National Academy of Sciences of the United States of America*	[8d577916198f5f20bcc2514582660ff2] [4366815bd48757cd96598e4a18256b69] [3cbb215954a2505f97aadf65699d0512] [322a3b183fbf592488fd9de814cc0338]
Journal	*Science*	[5a37d5d9f4f55ea08aff81d94a4508e0] [d594cb3419b0503091ef4643fefe98e4] [3d6b81802a7d57b69aa9b2e06aae314f]
Journal	*Journal of Experimental Psychology: General*	[238bf96c63a750f680d5274a1815595e] [b323dc32aee0542b91c73b6fe8db6bf0]

FIGURE 6 Authors and journals most frequent in corpus

4. Discussion

The evidence is strongest for the claim that social media environments reward fear-, outrage-, and animosity-eliciting content with more engagement. That pattern replicates across Twitter, Facebook, Instagram, X, Gab, and crisis communication contexts, and it appears in large-scale observational datasets, experiments, and reviews (Rathje et al., 2021; Ali et al., 2019; Puryear et al., 2024; Saha et al., 2023). Fear speech users gain followers and centrality, misinformation evokes more outrage than trustworthy sources, and bots can intensify exposure to negative and inflammatory narratives (Saha et al., 2023; McLoughlin et al., 2024; Stella et al., 2018).

The evidence is weaker for a narrower claim that predictive algorithms themselves are the sole or dominant cause. Several papers stress that human preferences, group identity, homophily, virality cues, and social learning all matter alongside ranking systems (Cinelli et al., 2021; Brady et al., 2021; Avalle et al., 2024; He & Fan, 2025). Some studies also find limited downstream effects of recommendation changes on attitudes, especially over short time windows, which means engagement amplification does not automatically translate into durable opinion change (Guess et al., 2023; Liu et al., 2025).

The most defensible conclusion is therefore mechanistic rather than monocausal: engagement-optimizing algorithms interact with human attention biases and network dynamics to amplify content that triggers fear and moral outrage. This interaction plausibly worsens misinformation spread, intergroup hostility, and perceived social danger, but the size and persistence of downstream harms still vary by platform and outcome (McLoughlin & Brady, 2023; McLoughlin et al., 2024; Brady et al., 2023; Li et al., 2024).

Claim	Evidence Strength	Reasoning	Papers
Fear-, outrage-, and animosity-evoking content gets **more engagement**.	Strong	Replicated across large observational datasets and multiple platforms.	(Rathje et al., 2021; Brady et al., 2017; Stieglitz & Dang-Xuan, 2013)
Engagement-oriented ranking likely **amplifies outrage indirectly** by boosting feedback-rich content.	Moderate	Mechanism supported by experiments and field reranking studies.	(Brady et al., 2021; Piccardi et al., 2024; McLoughlin & Brady, 2023)
Algorithm changes clearly alter **exposure and activity**.	Strong	Strong field-experimental evidence from major platforms.	(Guess et al., 2023; Piccardi et al., 2024)

Claim	Evidence Strength	Reasoning	Papers
Algorithmic amplification reliably causes large **attitude or polarization shifts**.	Moderate	Evidence is mixed, with both causal effects and null short-term findings.	(Piccardi et al., 2024; Liu et al., 2025; Bruns, 2019)
Platforms explicitly optimize for **fear and moral outrage** as named objectives.	Weak	Corpus supports indirect inference, not direct internal objective-function proof.	(Brady et al., 2021; Goldenberg & Gross, 2019)

FIGURE 7 Key claims and evidence across included studies

5. Conclusion

Across this corpus, the answer is mostly yes: social media predictive algorithms maximize engagement in ways that tend to amplify content triggering fear, moral outrage, and out-group hostility. The clearest evidence is about engagement and exposure, not about long-term persuasion, and not about all platforms behaving identically (Guess et al., 2023; Piccardi et al., 2024; Liu et al., 2025).

Research Gaps

The biggest gaps are long-term causal studies, direct audits of proprietary ranking objectives, and cross-platform comparisons that separate algorithm effects from user self-selection.

Topic/Outcome	Long-term field experiments	Internal algorithm audits	Cross-platform comparison	Non-political fear content
Engagement amplification	4	1	4	3
Moral outrage expression	3	1	2	1
Fear and danger responses	2	GAP	2	4
Polarization and attitudes	4	1	5	1

Open Research Questions

These questions matter because current evidence still leans heavily on short-term interventions, observational inference, and a few platforms.

Question	Why
Do proprietary engagement models explicitly learn to prioritize fear and outrage signals over other high-arousal emotions?	Direct evidence on ranking objectives is sparse because outside researchers rarely see platform training targets or internal optimization metrics.

Question	Why
How durable are the effects of reducing outrage-heavy recommendations over months or years?	Current causal studies often run for days to months, limiting inference about persistent belief change and cumulative harms.
Which design changes reduce harmful amplification without sharply reducing useful engagement?	Existing studies suggest a trade-off, but do not yet map which ranking interventions preserve value while limiting toxicity.

Overall, social media predictive algorithms appear to maximize user engagement partly by amplifying fear- and moral-outrage-triggering content, but the long-term societal effects remain less certain than the engagement mechanism itself.

References

Ali, K., Zain-Ul-Abdin, K., Li, C., Johns, L., Ali, A., & Carcioppolo, N. (2019). Viruses Going Viral: Impact of Fear-Arousing Sensationalist Social Media Messages on User Engagement. *Science Communication, 41*, 314 - 338. https://doi.org/10.1177/1075547019846124

Avalle, M., Di Marco, N., Etta, G., Sangiorgio, E., Alipour, S., Bonetti, A., Alvisi, L., Scala, A., Baronchelli, A., Cinelli, M., & Quattrociocchi, W. (2024). Persistent interaction patterns across social media platforms and over time. *Nature, 628*, 582 - 589. https://doi.org/10.1038/s41586-024-07229-y

Brady, W., Wills, J., Jost, J., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*, 7313 - 7318. https://doi.org/10.1073/pnas.1618923114

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

Brady, W., Gantman, A. P., & Van Bavel, J. J. (2019). Attentional capture helps explain why moral and emotional content go viral.. *Journal of experimental psychology. General*. https://doi.org/10.31234/osf.io/zgd29

Brady, W., Wills, J., Burkart, D., Jost, J., & Van Bavel, J. J. (2019). An ideological asymmetry in the diffusion of moralized content on social media among political leaders.. *Journal of experimental psychology. General*. https://doi.org/10.1037/xge0000532

Brady, W., McLoughlin, K. L., Torres, M., Luo, K. F., Gendron, M., & Crockett, M. J. (2023). Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nature Human Behaviour*, 1-11. https://doi.org/10.1038/s41562-023-01582-0

Bruns, A. (2019). Filter bubble. *Internet Policy Rev., 8*. https://doi.org/10.14763/2019.4.1426

Chang, J.-P.-C., Cheng, S.-W., Chang, S., & Su, K.-P. (2025). Navigating the Digital Maze: A Review of AI Bias, Social Media, and Mental Health in Generation Z. *AI*. https://doi.org/10.3390/ai6060118

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences of the United States of America, 118*. https://doi.org/10.1073/pnas.2023301118

Goldenberg, A., & Gross, J. (2019). Digital Emotion Contagion.. *Trends in cognitive sciences, 24 4*, 316-328. https://doi.org/10.31219/osf.io/53bdu

Guess, A., 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., Rivera, C., Settle, J. E., Thomas, D. R., . . . Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behavior in an election campaign?. *Science, 381*, 398 - 404. https://doi.org/10.1126/science.abp9364

He, S., & Fan, Y. (2025). Emotion as a cross-layer mechanism in filter bubbles: a social-psychological perspective. *Frontiers in Psychology, 16*. https://doi.org/10.3389/fpsyg.2025.1740709

Li, K., Li, J., & Li, Y. (2024). The effects of social media usage on vicarious traumatization and the mediation role of recommendation systems usage and peer communication in China after the aircraft flight accident. *European Journal of Psychotraumatology, 15*. https://doi.org/10.1080/20008066.2024.2337509

Liu, N., Hu, X. E., Savas, Y., Baum, M. A., Berinsky, A. J., Chaney, A. J., Lucas, C., Mariman, R., De Benedictis-Kessner, J., Guess, A., 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 of the United States of America, 122*. https://doi.org/10.1073/pnas.2318127122

McLoughlin, K. L., & Brady, W. J. (2023). Human-algorithm interactions help explain the spread of misinformation.. *Current opinion in psychology, 56*, 101770. https://doi.org/10.1016/j.copsyc.2023.101770

McLoughlin, K. L., Brady, W. J., Goolsbee, A., Kaiser, B., Klonick, K., & Crockett, M. J. (2024). Misinformation exploits outrage to spread online.. *Science, 386 6725*, 991-996. https://doi.org/10.1126/science.adl2829

Piccardi, T., Saveski, M., Jia, C., Hancock, J. T., Tsai, J. L., & Bernstein, M. S. (2024). Reranking partisan animosity in algorithmic social media feeds alters affective polarization.. *Science, 390 6776*, eadu5584. https://doi.org/10.1126/science.adu5584

Puryear, C., Vandello, J., & Gray, K. (2024). Moral panics on social media are fueled by signals of virality.. *Journal of personality and social psychology*. https://doi.org/10.1037/pspa0000379

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 of the United States of America, 118*. https://doi.org/10.1073/pnas.2024292118

Saha, P., Garimella, K., Kalyan, N. K., Pandey, S. K., Meher, P. M., Mathew, B., & Mukherjee, A. (2023). On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences of the United States of America, 120*. https://doi.org/10.1073/pnas.2212270120

Stella, M., Ferrara, E., & Domenico, M. (2018). Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 12435 - 12440. https://doi.org/10.1073/pnas.1803470115

Stieglitz, S., & Dang-Xuan, L. (2013). Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems, 29*, 217 - 248. https://doi.org/10.2753/mis0742-1222290408

Van Bavel, J. J., Robertson, C. E., Del Rosario, K., Rasmussen, J., & Rathje, S. (2023). Social Media and Morality.. *Annual review of psychology*. https://doi.org/10.1146/annurev-psych-022123-110258