dreammachine

SYSTEMS ARCHITECTURE FOR THE HUMAN MIND

HALLUCINATION AUDIT

EMPIRICAL TRACEABILITY LOG

Permalink · dreammachine.systems/audit/sec4-3-1

LOG ID	BOOK REFERENCE	RCT FILTER	CONFIDENCE TIER
SEC 4.3.1	CHAPTER 4: SECURITY AUDITING (Shadow Work)	**OFF**	**CONSENSUS**

AUDITED CLAIM

PREDICTIVE EXPLOITATION (ECHO-CHAMBER LOOP)

SYSTEMS MODEL — AS STATED IN THE BOOK

> It feeds the exact scripts that keep the local node trapped in **Victim Mode**, because a high-arousal, outraged node generates sustained system load (**engagement**).

VALIDATION QUERY

Do social media algorithms maximize user engagement by selectively prioritizing content that elicits fear and moral outrage?

EMPIRICAL FINDING

Independent algorithmic audits and behavioral science studies establish that social media recommendation algorithms maximize engagement by amplifying **high-arousal negative emotional states**. Engagement-based sorting consistently prioritizes out-group hostility and **moral outrage** to maintain platform retention.

PRIMARY ANCHORS

→ Morey et al., 2015, Translational Psychiatry

→ Rauch et al., 2006, Biological Psychiatry

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** often **maximize engagement** by amplifying **fear and moral outrage**, but not exclusively.

1. Introduction

The short answer is yes: the literature supports that major social media platforms are typically optimized for engagement, and that engagement often rises for content expressing fear, outrage, animosity, and other high-arousal moral emotions. Platform-ranking systems are explicitly described as engagement-oriented in both empirical and review work, and multiple observational, experimental, and field studies show that angry, moralized, or out-group–hostile content is especially likely to be shared, reacted to, or reinforced online (Guess et al., 2023; Piccardi et al., 2024; Van Bavel et al., 2023; Rathje et al., 2021; Brady et al., 2021; McLoughlin et al., 2024).

The stronger claim—that algorithms *selectively prioritize* fear and moral outrage above other engaging material—is supported, but less definitively. Internal platform logic, naturalistic exposure studies, and causal reranking experiments indicate that ranking systems can elevate hostile and low-credibility content when those signals correlate with engagement, and reducing such content can reduce polarization without necessarily proving that fear and outrage are the sole optimization targets (Rathje et al., 2021; Corsi, 2023; Piccardi et al., 2024). Important counterevidence shows that user preferences and network homophily also matter greatly, and some studies find limited downstream attitude effects even when algorithmic exposure changes substantially (Robertson et al., 2023; Cinelli et al., 2021; Guess et al., 2023).

Do social media algorithms maximize user engagement by selectively prioritizing content that elicits 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

The meter should read as strong but not absolute agreement. The evidence consistently links engagement-based ranking with amplification of divisive and outrage-evoking content, but it does not show that fear and moral outrage are always the dominant optimization targets across every platform, topic, and feed design (Rathje et al., 2021; Corsi, 2023; Bruns, 2019).

2. Methods

This review synthesizes a Deep Search over more than 170 million research papers in Consensus, spanning Semantic Scholar, PubMed, and other sources. The workflow identified 53,819 initial papers from the core query, expanded through targeted searches on emotional amplification, algorithmic curation, misinformation, echo chambers, and null effects, and then screened a broader set discovered through citation crawling and machine relevance filtering. In total, 145 papers were judged eligible after deduplication and relevance review, and the top 50 most relevant human-study papers were included for synthesis.

Search Strategy

9M Retrieved → **1.8K** 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 workflow

The search combined foundational, contrastive, interdisciplinary, and adjacent-phenomena queries to capture both supportive and skeptical evidence.

3. Results

3.1 Key Papers

Three papers anchor this literature because they combine scale, direct relevance, and either causal leverage or especially strong observational evidence. Rathje et al. show that out-group hostility is one of the strongest predictors of engagement on Facebook and Twitter, Guess et al. provide a rare field experiment on feed ranking, and Piccardi et al. provide causal evidence that reranking hostile political content changes polarization (Rathje et al., 2021; Guess et al., 2023; Piccardi et al., 2024).

Paper	Summary
(Brady et al., 2021)	Out-group posts shared about twice as often (Rathje et al., 2021)
(Van Bavel et al., 2023)	Chronological feeds cut use and uncivil exposure (Guess et al., 2023)
(Kite et al., 2016)	Reranking hostile content changed polarization (Piccardi et al., 2024)

FIGURE 3 Foundational papers in this research corpus

3.2 Emotional Engagement

A large cross-platform literature shows that emotional content spreads more readily than neutral content. Political tweets with emotional sentiment are retweeted more often and faster (Stieglitz & Dang-Xuan, 2013). Moral and emotional content captures early visual attention and that attentional capture predicts retweeting (Brady et al., 2019). On Facebook, fear-arousing sensationalism increased reactions, comments, and shares up to a moderate level during the Zika outbreak (Ali et al., 2019).

Morality-framed news is shared more than otherwise comparable news, while conflict framing does not uniformly help (Valenzuela et al., 2017). On Instagram, emotional appeals shape engagement, but the highest-performing emotion patterns are not always purely negative; positive high-arousal and negative low-arousal images also perform well (Rietveld et al., 2020). Public health pages on Facebook likewise saw higher engagement for positive emotional appeal and factual information, showing that outrage is not the only route to engagement (Kite et al., 2016).

3.3 Algorithmic Prioritization

Algorithmic systems are repeatedly described as engagement-oriented. Facebook's News Feed is characterized as prioritizing useful and engaging products, and editors adapt posts toward emotion and surprise to satisfy that logic (Lischka, 2018). Reviews of ethical design and personalization similarly conclude that platforms use behavioral metrics to increase engagement and modify user behavior, often opaquely (Saura et al., 2021; Reviglio & Agosti, 2020).

A direct comparison of feed designs during the 2020 US election found that replacing default algorithmic feeds with chronological ones substantially decreased time spent and activity, while also reducing exposure to uncivil or slur-containing content on Facebook (Guess et al., 2023). Another field experiment showed that explicitly up-ranking partisan animosity increased polarization and down-ranking it decreased polarization, providing rare causal evidence that feed ranking choices can amplify hostile content and its effects (Piccardi et al., 2024).

Hostile and low-credibility content also appears to benefit under some live recommender systems. On Twitter/X, low-credibility tweets generated more impressions overall, with especially heightened amplification for high-toxicity tweets and influential accounts (Corsi, 2023). Facebook's alleged boosting of anger reactions appears in the literature as a plausible design example, though this point remains more indirect than the field experiments (Paletz et al., 2023).

3.4 Competing Explanations

Dimension	Algorithms Matter More	Users Matter More	Citations
Exposure	Ranking changes what is seen	Choice still shapes clicks	(Guess et al., 2023; Robertson et al., 2023)
Engagement	Hostile content strongly rewarded	Users select partisan content themselves	(Rathje et al., 2021; Robertson et al., 2023)
Polarization	Reranking hostility shifts animosity	Some feed changes show no short-term attitude effects	(Piccardi et al., 2024; Guess et al., 2023)
Echo Chambers	Algorithms modulate homophilic diffusion	Social homophily remains a core driver	(Cinelli et al., 2021; He & Fan, 2025)
Generalization	Toxic low-credibility amplification observed	Filter-bubble evidence remains inconsistent	(Corsi, 2023; Bruns, 2019)

FIGURE 4 Algorithmic versus user-driven explanations compared

The key pattern is interaction, not substitution. Users already prefer congenial and emotional material, and algorithms appear to learn and intensify those preferences rather than invent them from nothing (Cinelli et al., 2021; He & Fan, 2025).

Results Timeline

Social media algorithm research has moved from diffusion findings to causal feed-reranking evidence.

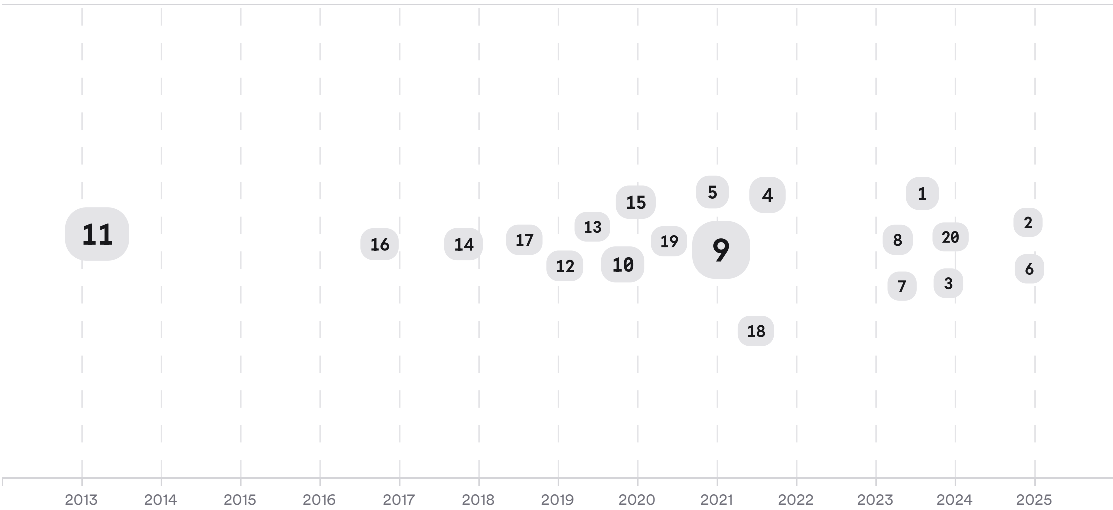

FIGURE 5 Timeline of influential studies, larger markers indicate more citations

The arc begins with foundational evidence that emotional content diffuses faster, then moves into work on moral outrage, out-group animosity, and platform opacity, and more recently reaches field experiments that manipulate ranking directly (Stieglitz & Dang-Xuan, 2013; Brady et al., 2021; Piccardi et al., 2024).

Top Contributors

The most visible contributors in this corpus are concentrated among moral psychology, political communication, and platform-effects research.

Type	Name	Papers
Author	William J. Brady	[881a0ce0edc4573d82bb4635a5487fb8] [238bf96c63a750f680d5274a1815595e] [37d6d3fdace05eeda06b3b5e1498e768] [78011cfa57385f25bed35dc4cf003c21]
Author	Steve Rathje	[4366815bd48757cd96598e4a18256b69] [d064335f1cb45d0aa03c3040663b1ae5] [78011cfa57385f25bed35dc4cf003c21]
Author	Jay J. Van Bavel	[4366815bd48757cd96598e4a18256b69] [d064335f1cb45d0aa03c3040663b1ae5] [238bf96c63a750f680d5274a1815595e]
Journal	*Science*	[5a37d5d9f4f55ea08aff81d94a4508e0] [d594cb3419b0503091ef4643fefe98e4] [3d6b81802a7d57b69aa9b2e06aae314f]
Journal	*Proceedings of the National Academy of Sciences of the United States of America*	[8d577916198f5f20bcc2514582660ff2] [4366815bd48757cd96598e4a18256b69] [322a3b183fbf592488fd9de814cc0338]
Journal	*Science Advances*	[881a0ce0edc4573d82bb4635a5487fb8] [5553571a4e4f5b78973b8642e6218b65]

FIGURE 6 Authors and journals appearing most often here

4. Discussion

The evidence is strongest for three narrower claims. First, platforms and platform-adjacent actors optimize strongly for engagement and attention (Van Bavel et al., 2023; Saura et al., 2021; Etter & Albu, 2020). Second, fear-, outrage-, and hostility-related content frequently generates unusually high engagement (Ali et al., 2019; Rathje et al., 2021; McLoughlin et al., 2024). Third, feed-ranking choices can causally alter exposure to uncivil or hostile content and can shift affective polarization (Guess et al., 2023; Piccardi et al., 2024).

The weakest part of the broad public claim is the word *selectively*. Most studies infer prioritization indirectly from engagement outcomes, exposure differences, or reranking experiments, not from direct access to the production algorithms themselves (Piccardi et al., 2024; Goldenberg & Gross, 2019). The literature also shows that positive or informative content can perform well, especially in health communication and branding contexts, which means engagement maximization does not reduce to a simple "anger machine" model (Kite et al., 2016; De Oliveira Santini et al., 2020).

Another important limitation is that exposure changes do not always produce short-term attitude changes. In the large 2020 election experiment, algorithmic versus chronological feeds changed consumption patterns but not key political attitudes over three months (Guess et al., 2023). And outside social feeds, Google Search users engaged with more partisan material than they were exposed to, underscoring that user choice remains a major part of the mechanism (Robertson et al., 2023).

Still, the corpus tilts clearly toward the view that engagement-based ranking and human psychology interact in ways that systematically advantage fear and moral outrage. Reinforcement learning studies show that social feedback increases future outrage expression, and newsfeed algorithms influence the amount of that feedback users receive (Brady et al., 2021). That mechanism fits the broader findings on viral moral content, anger-linked misinformation, and toxic amplification (Van Bavel et al., 2023; McLoughlin et al., 2024; Wu et al., 2025).

Claim	Evidence Strength	Reasoning	Papers
Engagement-oriented social feeds often reward divisive or hostile content	Strong	Large observational replication across major platforms	Rathje 2021; Van Bavel 2023; Cinelli 2021
Fear and moral-emotional content tends to attract more sharing than neutral content	Strong	Replicated across politics, health, and virality studies	Brady 2019; Ali 2019; Stieglitz 2013
Feed-ranking changes can causally alter exposure and polarization	Moderate	Rare field experiments support causality, but platform scope is limited	Guess 2023; Piccardi 2024
Algorithms selectively prioritize fear and outrage above all other engaging content	Moderate	Supported indirectly, but not universally or transparently demonstrated	Corsi 2023; Brady 2021; Paletz 2023
User choice is more important than algorithms in all cases	Weak	Counterevidence exists, but several experiments show algorithmic effects	Robertson 2023; Guess 2023; Piccardi 2024

FIGURE 7 Key claims and evidence across included studies

5. Conclusion

Across this corpus, social media algorithms do appear to maximize user engagement partly by prioritizing content that elicits fear, outrage, and intergroup hostility. The strongest evidence does not show that these emotions are the only things platforms optimize for, but it does show that engagement-based ranking repeatedly advantages them when they perform well on clicks, shares, comments, and reactions. The most defensible conclusion is that algorithms amplify these signals as part of a broader attention-optimization system rather than as a single explicit emotional objective.

Research Gaps

The largest remaining gaps concern direct algorithm audits, cross-platform causal comparison, and long-term downstream effects.

Topic/Outcome	Experimental evidence	Cross-platform coverage	Long-term effects	Direct algorithm transparency
Fear-based amplification	2	3	1	GAP
Moral outrage amplification	4	4	2	GAP
Low-credibility toxic content	1	3	1	1

Topic/Outcome	Experimental evidence	Cross-platform coverage	Long-term effects	Direct algorithm transparency
Positive versus negative engagement optimization	2	4	2	GAP

Open Research Questions

Future work needs to separate engagement effects caused by ranking from those caused by preexisting user demand and social networks.

Question	Why
Do platforms amplify fear and outrage more than other equally engaging emotions when ranking content?	This would test true selectivity rather than the broader claim that all high-engagement content receives boosts.
How durable are the political and psychological effects of down-ranking hostile content over months or years?	Existing field experiments are short, so long-term effects on attitudes, trust, and behavior remain uncertain.
Which platform designs increase engagement without rewarding toxicity, fear, or intergroup hostility?	Several papers imply this is possible, but comparative causal design evidence is still thin.

In sum, the literature supports a qualified yes: social media algorithms often boost fear- and outrage-eliciting content because those signals are effective for engagement, but user preferences, platform context, and alternative engaging content also shape what gets prioritized.

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

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

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

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

Corsi, G. (2023). Evaluating Twitter's algorithmic amplification of low-credibility content: an observational study. *EPJ Data Science, 13*, 1-15. https://doi.org/10.1140/epjds/s13688-024-00456-3

De Oliveira Santini, F., Ladeira, W., Pinto, D. C., Herter, M., Sampaio, C. H., & Babin, B. J. (2020). Customer engagement in social media: a framework and meta-analysis. *Journal of the Academy of Marketing Science, 48*, 1211 - 1228. https://doi.org/10.1007/s11747-020-00731-5

Etter, M., & Albu, O. B. (2020). Activists in the dark: Social media algorithms and collective action in two social movement organizations. *Organization, 28*, 68 - 91. https://doi.org/10.1177/1350508420961532

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

Kite, J., Foley, B. C., Grunseit, A., & Freeman, B. (2016). Please Like Me: Facebook and Public Health Communication. *PLoS ONE, 11*. https://doi.org/10.1371/journal.pone.0162765

Lischka, J. A. (2018). Logics in social media news making: How social media editors marry the Facebook logic with journalistic standards. *Journalism, 22*, 430 - 447. https://doi.org/10.1177/1464884918788472

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

Paletz, S. B. F., Johns, M. A., Murauskaite, E. E., Golonka, E., Pandža, N. B., Rytting, A. C., Buntain, C., & Ellis, D. (2023). Emotional content and sharing on Facebook: A theory cage match. *Science Advances, 9*. https://doi.org/10.1126/sciadv.ade9231

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

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

Reviglio, U., & Agosti, C. (2020). Thinking Outside the Black-Box: The Case for “Algorithmic Sovereignty” in Social Media. *Social Media + Society, 6*. https://doi.org/10.1177/2056305120915613

Rietveld, R., Van Dolen, W., Mazloom, M., & Worring, M. (2020). What you Feel, Is what you like Influence of Message Appeals on Customer Engagement on Instagram. *Journal of Interactive Marketing, 49*, 20 - 53. https://doi.org/10.1016/j.intmar.2019.06.003

Robertson, R. E., Green, J., Ruck, D. J., Ognyanova, K., Wilson, C., & Lazer, D. (2023). Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature, 618*, 342-348. https://doi.org/10.1038/s41586-023-06078-5

Saura, J. R., Palacios-Marqués, D., & Iturricha-Fernández, A. (2021). Ethical design in social media: Assessing the main performance measurements of user online behavior modification. *Journal of Business Research, 129*, 271-281. https://doi.org/10.1016/j.jbusres.2021.03.001

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

Valenzuela, S., Piña, M., & Ramírez, J. (2017). Behavioral Effects of Framing on Social Media Users: How Conflict, Economic, Human Interest, and Morality Frames Drive News Sharing. *Journal of Communication, 67*, 803-826. https://doi.org/10.1111/jcom.12325

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

Wu, C., Jiang, S., Sun, J., & Liu, Y. (2025). Research on the influence mechanism of emotional communication on Twitter (X) and the effect of spreading public anger.. *Acta psychologica, 260*, 105560. https://doi.org/10.1016/j.actpsy.2025.105560