

Safety Without Surveillance™ • Education Without Exploitation

Investigating algorithmic systems to protect children and families

The Dark Side of Algorithms: An Investigative Report on X (formerly Twitter)

Published by: My AI Helper Series | Digital2Digital Publishing

Website: MyAIHelperSeries.com

Report Date: February 2026

Executive Summary

This investigative report examines the algorithmic systems operating on X (formerly Twitter) under Elon Musk's ownership, revealing significant evidence of systemic biases, content moderation failures, and business model incentives that amplify harmful content. Drawing from peer-reviewed academic studies, parliamentary investigations, regulatory findings, and documented real-world harms, this report demonstrates how X's algorithmic architecture contributes to political polarization, misinformation proliferation, and societal destabilization.

Key Findings:

- **Algorithmic Political Bias:** Multiple independent audits confirm X's algorithm exhibits measurable right-leaning political bias, particularly in out-of-network recommendations for new users, with right-leaning accounts experiencing the highest levels of exposure inequality.
- **Engagement-Driven Harm:** X's monetization model, which pays creators based on engagement from Premium subscribers, creates perverse incentives that reward inflammatory, divisive, and misleading content—often without regard for veracity or societal harm.
- **Content Moderation Collapse:** Following Musk's acquisition, X dismantled professional content moderation teams, replacing them with the failing Community Notes system. Hate speech reports increased **50-273%**, while enforcement actions against harmful content plummeted by over **97%**.
- **Transparency as Illusion:** Despite claims of "open-sourcing" the algorithm, X has actually increased opacity by transitioning to AI-driven ranking systems (Grok-based) while simultaneously restricting researcher access to data and eliminating meaningful transparency reports.
- **Real-World Violence:** The platform's algorithms played a documented role in amplifying disinformation that contributed to the 2024 Southport riots in the UK, where false claims achieved **155 million impressions** and directly fueled anti-Muslim violence.
- **Regulatory Sanctions:** The European Union has formally fined X **€120 million (\$140 million)** for violations of the Digital Services Act, citing deceptive design practices, inadequate advertising transparency, and systematic obstruction of research access.

1. Algorithmic Architecture: Engineering Bias into the System

1.1 The Shift to AI-Driven Opacity

In early 2026, X released what it termed an "open-source" version of its recommendation algorithm. However, researchers immediately identified this as a transparency failure rather than a success. Unlike the 2023 release, which included specific weighting factors (e.g., replies valued at 27 retweets, author responses worth 75 retweets), the current system relies on a Grok-based large language model to rank posts—a shift that makes the algorithm significantly more opaque.

"So much more of the decisionmaking ... is happening within black box neural networks that they're training on their data... More and more of the decisionmaking power of these algorithms is shifting not just out of public view, but actually really out of view or understanding of even the internal engineers that are working on these systems."

— Professor John Thickstun, Cornell University

The code release excluded critical information about training data, model weights, and ranking methodologies—elements essential for meaningful external audit. As one researcher observed, "They can't possibly draw those conclusions from what was released."

1.2 Documented Political Bias in Algorithmic Curation

The 2024 Election Audit Findings

A comprehensive six-week audit conducted during the 2024 U.S. Presidential Election by researchers at USC's Viterbi School of Engineering deployed 120 sock-puppet monitoring accounts to capture algorithmic recommendations. The study, published in the Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, revealed systematic biases:

- Approximately **50%** of tweets in X's user timelines are personalized recommendations from accounts users do not follow
- Right-leaning accounts experience the highest level of exposure inequality (Gini coefficient analysis, $p < 0.001$)
- Neutral accounts (new users) show a default right-leaning bias in content exposure
- Both left- and right-leaning accounts encounter amplified exposure to ideologically aligned users and reduced exposure to opposing viewpoints
- The algorithm amplifies political commentators and influencers over traditional institutional voices

The study found that among the top 20 recommended users for neutral accounts, right-leaning users comprised **30.16%** of exposure compared to only **12.92%** for left-leaning users—a statistically significant disparity that suggests the algorithm's default state favors conservative content.

The Musk Amplification Effect

Separate research by Timothy Graham (Queensland University of Technology) and Mark Andrejevic (Monash University) identified a "change point" in engagement metrics coinciding precisely with Elon Musk's endorsement of Donald Trump on July 13, 2024. Using CUSUM (Cumulative Sum) statistical analysis, they found:

- Musk's posts received a boost of approximately **6.4 million views** post-endorsement
- This represented a **138.27% increase** in views independent of general platform trends
- The engagement surge was unique to Musk's account compared to other high-profile accounts

The researchers concluded this suggests "the possibility of algorithmic prioritization or bias, positioning Musk's content favorably in terms of platform visibility and user engagement."

2. The Monetization of Misinformation: Engagement at All Costs

2.1 The Shift to Engagement-Based Revenue

In October 2024, X fundamentally altered its creator monetization model, shifting from ad-revenue sharing to direct payments based on engagement from X Premium subscribers. This change has profound implications for content incentives:

- Creators earn based on likes, replies, reposts, and bookmarks from Premium subscribers only
- Non-Premium engagement counts "very little toward payout"
- No demonetization policies for misinformation (unlike YouTube, Meta)
- Requirements: Premium subscription, 500 followers, 5 million impressions in 3 months

"Often, when platforms begin to reward engagement, creators will begin posting anything that drives a discussion of some sort, including posts that are designed to enrage users, forcing them to reply or comment... X, however, is going in the opposite direction with its new policy."

— TechCrunch

2.2 Research on Harmful Content Amplification

A study published in EPJ Data Science (2024) evaluated Twitter's algorithmic amplification of low-credibility content using impressions data. Key findings:

- Tweets containing URLs from low-credibility domains achieved **+19.2%** higher visibility for COVID-19 content
- **+95.8%** higher visibility for climate change misinformation
- High-toxicity tweets exhibited heightened amplification when containing low-credibility content
- The effect was driven primarily by high-engagement, high-follower accounts

3. Content Moderation: The Systematic Dismantling of Safety

3.1 The Collapse of Professional Moderation

Following Musk's acquisition, X implemented dramatic changes to content moderation:

- **Staff Reductions:** Trust and Safety teams decimated through layoffs
- **Policy Rollbacks:** Rules on COVID-19 misinformation eliminated; misgendering and deadnaming no longer classified as hate speech
- **Account Reinstatements:** Previously banned extremist accounts welcomed back to platform
- **Automated Failures:** Lucy Connolly case (wife of Conservative councillor) posted explicit calls for attacks on migrant hotels; X's automated systems deemed it acceptable despite violating UK law on incitement to racial hatred

UK Parliament findings documented that hate speech on X rose **50-273%** in months following Musk's takeover.

3.2 The Community Notes Failure

X's Community Notes system, touted as a "democratic" alternative to professional moderation, has proven systematically ineffective:

- **74%** of accurate Community Notes on US election misinformation never displayed to users
- Posts without Community Notes promoting false political narratives garnered **billions of views**
- Fact-checked content was outpaced by misinformation reach at a ratio of **13:1**
- The system requires consensus, which is rarely achieved on polarizing issues—precisely where fact-checking is most needed

ProPublica/Columbia University Investigation:

- Examined 200 distinct misleading claims fact-checked by independent sources
- Identified 2,000 tweets by verified accounts perpetuating these claims
- Collective views: **half a billion impressions**
- **80%** of debunked posts had no Community Note attached
- More than 80 debunked claims were never clarified with a note

The Washington Post confirmed that when notes do appear, they are **four times more likely** to be approved on posts from Republican politicians than Democrats.

3.3 Enforcement Statistics Collapse

Comparative analysis of X's transparency data reveals the scale of moderation decline:

- **2021 (H2):** 104,565 accounts suspended for hateful conduct
- **2024:** 2,361 accounts suspended for hateful conduct
- **Reduction:** **97.7% decrease** in enforcement actions

While X attributes this to changing measurement methods, the shift coincides with policy rollbacks and staff reductions rather than improved platform safety.

4. Case Study: The Southport Riots—Algorithms to Real-World Violence

The July-August 2024 riots in the UK provide a documented case study of how X's algorithms translate online disinformation into offline violence.

4.1 The Disinformation Proliferation

Following the stabbing attack in Southport that killed three young girls, false claims about the perpetrator's identity (Muslim, asylum seeker, arrived by boat) spread rapidly across X:

- False claims achieved **155 million impressions** on X between July 29 and August 9, 2024
- The false name was seen **420,000 times** with potential reach of **1.7 billion people**
- Anti-migrant content received **31.1 million impressions** in the same period
- Posts with keyword "Muslim" increased **242%**

4.2 Algorithmic Amplification of Conspiracy Theories

Research funded by the London School of Economics analyzed posts from two verified X accounts during the riots period and found:

1. Amplification of Racist Conspiracy Theories:

Posts featuring visual representations of racist conspiracy theories ("White genocide," "The Great Replacement") were amplified approximately **30% more** than other content. These theories, which claim deliberate replacement of White populations through immigration, have historically motivated violent extremism.

2. AI-Generated Disinformation:

- 39 posts from one account used Generative AI to create racist imagery
- Average views per AI-generated post were nearly **three times higher** than other content
- One post alone obtained **11 million views**
- Images depicted racialized Muslim men as sexual predators targeting White British girls

"Social media companies' responses were inconsistent and inadequate, often enabling, if not encouraging, this viral spread, with evidence that they may have profited due to the heightened engagement."

— UK Parliament's Science and Technology Committee

5. Transparency and Research Access: Obstruction and Obfuscation

5.1 The Illusion of Open Source

While X has released portions of its recommendation algorithm code, researchers emphasize this provides limited meaningful transparency. Critical gaps include:

- No information about training data used for Grok-based ranking models
- Exclusion of weighting factors and scoring methodologies (redacted "for security reasons")
- No visibility into feature engineering or model architecture
- Inability to replicate or audit actual platform behavior

5.2 Systematic Obstruction of Research

The European Commission's December 2025 decision documented X's systematic efforts to prevent legitimate research:

- Free Basic API limited to 1,500 posts per month (insufficient for research)
- Commercial API costs up to **\$5,000 per month**
- Dedicated research application form launched November 2023
- Not a single application approved before January 26, 2024
- **95.8% rejection rate** as of May 2024
- Successful applicants assigned to \$5,000 tier with 1 million tweets/month limit—a tenfold reduction from Twitter's previous academic program

6. Regulatory Response: The European Union's Digital Services Act

6.1 The €120 Million Fine

On December 5, 2025, the European Commission issued its first non-compliance decision under the Digital Services Act (DSA), fining X **€120 million (\$140 million)** for three categories of violations:

Violation 1: Deceptive Design (Blue Checkmarks)

- X "materially changed the verification process as compared to cross-industry standards"
- Maintained visual design implying authenticity while distributing verification to anonymous paying subscribers
- Human review averaged only **53-79 seconds** per account—insufficient for meaningful identity verification
- Premium subscribers receive "reply prioritization" allowing users to "pay for reach"

Violation 2: Advertising Repository Deficiencies

- Database undermined by "design features and access barriers"
- "Excessive delays in processing" researcher requests
- **42%** of ads not retrievable in repository

Violation 3: Researcher Access Barriers

- "Unnecessary barriers" for researchers accessing public data
- **95.8%** application rejection rate
- Stymies research into systemic risks facing European users

10. Conclusions and Recommendations

10.1 Summary of Evidence

The evidence presented in this report demonstrates that X's algorithmic systems, business model, and governance structure create systematic harms:

1. **Political Manipulation:** Measurable algorithmic bias favoring right-leaning content, particularly for new users, with documented amplification of the platform owner's political content coinciding with electoral endorsements.
2. **Misinformation Proliferation:** Engagement-based monetization rewards false and harmful content; Community Notes fails to scale; professional moderation has been dismantled.
3. **Real-World Violence:** Documented role in amplifying disinformation that fueled the Southport riots, with false claims achieving 155 million impressions and platform systems deeming calls for violence "acceptable."
4. **Transparency Theater:** Claims of open-sourcing algorithms mask increased opacity through AI-driven ranking; systematic obstruction of legitimate research.
5. **Regulatory Sanctions:** Formal findings of DSA violations by the European Union, with €120 million fine for deceptive design, inadequate transparency, and researcher access obstruction.

10.3 Final Assessment

X under Elon Musk represents a case study in what happens when algorithmic systems are deployed with minimal oversight, transparency, or accountability. The platform's shift toward engagement-based monetization, combined with the dismantling of content moderation infrastructure and documented political manipulation, creates a dangerous environment where harmful content is systematically amplified for profit.

The Southport riots demonstrate that these are not abstract concerns—the algorithmic amplification of disinformation on X directly contributed to real-world violence against minority communities. The European Union's DSA enforcement action shows that regulatory intervention is both necessary and possible.

As the research in Science demonstrated, algorithmic curation of hostile content can shift political attitudes in one week by an amount that normally takes three years. The stakes could not be higher. X's algorithms are not neutral arbiters of information; they are active participants in shaping political discourse, social cohesion, and democratic integrity.

The "dark side of algorithms" is not a bug—it is the inevitable result of systems optimized for engagement above all other values. Until platforms are held accountable for the societal harms their algorithms produce, the dark side will continue to grow.

References

Note: All references from the original report are maintained. This is an abbreviated list for the PDF version.

- Algorithmic Effects on Account Visibility in Pre-X Twitter. arXiv:2512.06129
- Ye, J., Luceri, L., & Ferrara, E. (2025). Auditing Political Exposure Bias: Algorithmic Amplification on Twitter/X During the 2024 U.S. Presidential Election
- Graham, T., & Andrejevic, M. (2024). A computational analysis of potential algorithmic bias on platform X during the 2024 US election
- Jia, C., et al. (2025). Algorithmic influence on political polarization. Science
- Corsi, G. (2024). Evaluating Twitter's algorithmic amplification of low-credibility content. EPJ Data Science
- Jones, M. O. (2025). Written evidence submitted to UK Parliament Science and Technology Committee
- Center for Countering Digital Hate. (2024). How X's Community Notes System Falls Short
- UK Parliament Science and Technology Committee. (2024). Social media, misinformation and harmful algorithms
- European Commission. (2025). Digital Services Act Decision regarding X
- Additional references available in full report at MyAIHelperSeries.com

MY AI HELPER SERIES

Safety Without Surveillance™

MyAIHelperSeries.com | Digital2Digital Publishing

© 2026 Digital2Digital Publishing | All Rights Reserved