Written evidence submitted by Marc Owen Jones (PhD) (SMH0071)

Associate Professor at Northwestern University in Qatar

**Unless otherwise indicated, statistics quoted here are generated from the researcher's own work and analysis on the riots*

1. The Southport riots of July-August 2024, which erupted following false claims about a stabbing incident, provide compelling evidence of how social media platforms contribute to real-world violence. This submission analyses comprehensive data from X (formerly Twitter) and other platforms, documenting how business models, personal ideology, algorithms, and content moderation failures enabled the rapid spread of xenophobic disinformation that preceded violent attacks on minorities and migrant housing. The rise of 'disinfluencers' (routine spreaders of false information, often with disproportionate ability to influence trends) should also be noted as an important aspect of the post-truth economy.

2. It is imperative to mention from the outset that there is not sufficient transparency of platforms and how they modulate and moderate content to allow for suitably informed. This privacy acts as a barrier against accountability and facilitates unaccountable disinformation. Expecting the private sector to regulate the public square is a cornerstone of the issue of healthy public discourse.

Southport riots preceded disinformation preceding months of anti-Muslim disinfo on META and other platforms

3. The killings of Southport did not happen in a vacuum. Firstly, it is important to recognise that during any terror-like attack, minorities are often reflexively blamed. Certainly, while anti-immigrant and anti-Muslim sentiment is not new, pertinent to this inquiry is social media. In the six months leading up to Southport, the UK was bombarded by a virulent anti-Muslim and anti-immigrant advertising campaign - mostly on Meta's platforms.[1] These narratives appeared to be targeting and riling up xenophobic sentiment ahead of European elections. [2]

---

[1] https://www.france24.com/en/live-news/20240708-shadow-campaign-global-influence-op-targets-qatar-in-wartime

[2] https://www.aljazeera.com/opinions/2024/7/24/the-qatar-plot-how-a-covert-influence-campaign-helped-europes-far-right

4. On Facebook alone, at least $1.2 million was spent on anti-immigrant advertising targeting UK, US, and European users in the months preceding the riots, suggesting systematic exploitation of platform advertising systems to promote harmful narratives. The campaign was not ‘attributed’ by Meta, meaning they do not know who was behind it. This raises the alarming point that they are receiving money from entities to engage in hate speech and disinfo, without even knowing who those entities are. [3] In terms of policy, there needs to be better mechanisms to ensure social media companies know who they are receiving money from.

5. A similar case where social media platforms were weaponized by a Tel Avis-based PR firm contracted by Israel’s Ministry of Diaspora Affairs and Combating Antisemitism. The operation targeted mostly English-speaking audiences with AI-generated content promoting anti-muslim sentiment and fears of Muslim immigration.[4]

6. The Southport case is particularly relevant to this inquiry because it demonstrates the full lifecycle of harmful content: from initial false claims, through algorithmic amplification, to real-world violence. Between 29th July and August 9th August, disinformation about the attacker's identity spread rapidly across social platforms, achieving at least 155 million impressions on X alone. This false narrative was subsequently linked to attacks on a mosque in Southport and migrant housing, demonstrating direct connections between online content and offline harm.

7. The platform mechanics on X that likely enabled this harmful content operated through six key mechanisms:

    - Paid Verification System: X replaced identity verification with a paid "blue tick" system, algorithmically boosting subscriber content regardless of accuracy.
    - Reduced Content Moderation: Significant cuts to moderation teams left harmful content unchecked, possibly influenced by ideological factors.
    - Monetization Incentives for Misinfo/disinfo: Engagement-driven monetization rewarded content regardless of its veracity and no de-monetization for disinfo
    - Algorithmic Amplification: Algorithms may have systematically prioritized inflammatory and divisive content over factual corrections.
    - Anonymity and Bots: The presence of anonymous accounts and bots contributed to the spread of disinformation.
    - Disinfluential Amplification: High-profile accounts, including Elon Musk’s, played a role in amplifying disinformation and inflammatory content.

---

[3] https://scontent.fdoh6-1.fna.fbcdn.net/v/t39.8562-6/455158590_1227906161699704_7318728570685925077_n.pdf?_nc_cat=100&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=BSu7-ZsPOrMQ7kNvgGTMWIF&_nc_oc=Adihcqay3q9vNQ4lPSA8vS1P6srJRh2433BHGmce7MIXW11WSAnEOGeizJLqStxeA6y5WIRgnYovaOq5hrSz-JkB&_nc_zt=14&_nc_ht=scontent.fdoh6-1.fna&_nc_gid=AH7BhSl2P-Ww_rc93FWvgKI&oh=00_AYC4Nb9-CDj-zsuV5N4I6PriBQB9fUVrC2z_TBNzL8aKlQ&oe=67C1531F
[4] https://cyberscoop.com/israel-influence-operations-stoic/

8. Throughout 2024, X underwent significant algorithmic and policy changes that reshaped the platform into a more polarized and engagement-driven space. One of the most notable shifts was the transformation of its verification and monetization systems. Users were now required to pay for X Premium (formerly Twitter Blue) to receive a blue checkmark, marking a departure from the previous system where verification was granted based on authenticity.[5]

9. In October 2024, X changed its revenue-sharing model, moving away from ad-based monetization and instead rewarding engagement. This meant that users earned money based on the number of likes, shares, and comments their posts received—provided those interactions came from other Premium users. Unlike other major platforms, X did not introduce policies to de-monetize or suspend accounts for posting misinformation, making engagement-driven incentives a powerful force in shaping content.[6]

10. Experts and observers noted a shift in the types of posts being amplified by the platform's algorithm. Many reported that divisive and inflammatory content received greater visibility, particularly posts supporting Donald Trump or attacking his political opponents. Former disinformation governance experts suggested that X's algorithms had been reconfigured to privilege misleading and polarizing rhetoric, leading to a notable increase in sensationalist content. [7]

11. Additionally, the platform introduced several controversial changes that affected user interactions. Likes were made private, meaning users could no longer see what others had liked, reducing transparency. Blocking was also altered, allowing blocked accounts to continue viewing public posts, effectively weakening users' ability to shield themselves from harassment. These adjustments raised concerns about user safety, particularly for those targeted by online abuse.[8]

12. The reduction in content moderation further contributed to the spread of harmful content. Not only at X, but at Meta, Discord, and other platforms[9]. Large-scale layoffs in moderation teams meant that misinformation and incendiary posts were left largely unchecked. Despite past commitments to tackling manipulated media, X lacked the safeguards other platforms maintained to curb disinformation and abuse. This regulatory vacuum enabled the rapid rise of influential but controversial accounts such as "Inevitable West," which quickly amassed a large following by posting divisive content. Musk himself interacted with and amplified such accounts, reinforcing their reach.[10]

---

[5] https://www.bbc.com/news/articles/c1elddq34p7o
[6] https://www.bbc.com/news/articles/c1elddq34p7o
[7] https://www.bbc.com/news/articles/c1elddq34p7o
[8] https://www.npr.org/2024/06/13/nx-s1-5004515/x-likes-hide-users-elon-musk
[9] https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html
[10] https://www.msn.com/en-gb/money/other/elon-musk-keeps-reposting-inevitable-west-s-attacks-

13. The Lucy Connolly case provides stark evidence of platform failures to prevent incitement to violence. Connolly, the wife of a Conservative councillor, posted explicit calls for attacks on migrant hotels following the spread of false information about the Southport stabbings. Despite this content clearly violating both platform rules and UK law regarding incitement to racial hatred, X's automated systems deemed it acceptable.[11]

14. Perhaps the most consequential shift was Elon Musk's growing political alignment with Donald Trump. In July 2024, Musk publicly endorsed Trump's presidential campaign. Later in the year, he was offered a government advisory position leading the Department of Government Efficiency (Doge), a move that further blurred the lines between his role as a tech CEO and a political influencer. This development suggested that X's policies could be increasingly shaped by political considerations, especially as Trump returned to power.[12]

15. These changes manifested in the Southport riots were characterised by the privileging of disinformation and hate speech. Posts spreading false or hateful information about Muslims and immigrants amassed 65% of total impressions related to the riots, while factual corrections and defences of targeted communities achieved only 13%. This disparity can't be said to be accidental but as a result directly of platform business models that prioritize engagement over accuracy, veracity and harm. It also points to a potentially increased dominance of xenophobic voices on X after Musk's takeover.

16. The amplification of harmful content followed clear patterns visible in the impression data. Anti-Muslim content achieved 80.3 million impressions and anti-immigrant content 31.1 million impressions, while neutral content reached 50.2 million and factual corrections only 31.7 million impressions. This disparity demonstrates an alarming amplification of disinformation and hate speech.

17. Appropriate interventions to attenuate false harmful information were not present or clearly evident. This was demonstrated by Musk's curation of hate speech and false information. Even without knowing the explicit details of algorithms, X was effectively rendered a disinformation delivery system during the riots - where hate speech and falsehoods outstripped truth.

18. Cross-platform analysis reveals this was not isolated to X. Evidence shows coordinated disinformation campaigns operated across multiple platforms, including Facebook, Telegram, and TikTok. However, X has been notable due to its role as a real-time source of breaking news.

---

on-keir-starmer-but-is-an-indian-fraudster-behind-the-account/ar-AA1yNOPa

[11] https://www.bbc.com/news/articles/cn8ljjjmpg5o

[12] https://www.bbc.com/news/articles/c1elddq34p7o

Disinfluencers

19. Social media algorithms likely played a decisive role in amplifying harmful content during the riots. Certain accounts repeatedly engage in false and xenophobic or hateful narratives - these so-called ‘dis/dysinfluencers’ include quasi-news accounts like visegrad24, who often hastily report in breaking news that the attacker was a Muslim or immigrant (before there is evidence).

20. A component of Algorithm manipulation involves highly followed and influential accounts amplifying false content. This directly impacts what is seen, and how popular it is. That the most followed person on X (Elon Musk) was actively promoting disinfluencer accounts during the riots equates to algorithm manipulation in ‘truth markets’. A retweet or engagement on a topic from Musk will significantly boost the visibility of content.

21. On July 29th, between 16:00 and 18:00 BST, a surge of tweets emerged spreading disinformation that the killer in Southport was "Ali Al Shakati" - an Arab/Muslim name. The account @artemisfornow, operated by Bernadette Spofforth, was the first publicly documented source of this fabricated name. The initial tweet falsely claimed that this supposed individual was both on an MI6 watchlist and known to Liverpool mental health services. The name appears to have been deliberately constructed to sound Muslim.[13]

22. The timeline of the Southport riots demonstrates the potential of causation between online disinformation and offline violence. Within hours of the initial incident, a fabricated narrative about "Ali Al Shakati" - a completely invented name designed to suggest Muslim involvement - achieved 27 million impressions. Shortly afterwards, far-right groups attacked the Southport mosque, with participants explicitly referencing the false social media claims as justification. The name ‘Ali Al Shakati’ gave legitimacy or credibility to the extant rumour that the attacker was a Muslim. Such ‘scaffolding’ of narratives enables them to resemble breaking news, lending them further believability.

23. Spofforth's tweet gained significant traction, receiving over a million impressions. This reach was particularly notable given Spofforth's background of promoting COVID-19 conspiracy theories and having previously had her Twitter account suspended.[14] The disinformation was further amplified by Channel3NowNews, a sophisticated-looking but fraudulent news operation. This account maintained the appearance of legitimacy through professional branding and a presence across multiple platforms including a website, YouTube, and Facebook pages. The operation was later traced to Farhan Asif in Pakistan, who was using these platforms to generate revenue through clickbait and marketing techniques.

---

[13] https://www.bbc.com/news/articles/crl8nwx6ynzo
[14] https://academic.oup.com/isq/article/68/4/sqae131/7815709

24. Both individuals faced legal consequences for their actions. Cheshire Police arrested Spofforth, though she was later released on bail without charges. Similarly, Pakistani authorities arrested and subsequently released Farhan Asif. The arrests sparked considerable online controversy, particularly among right-wing groups. The Free Speech Union, led by Toby Young, notably championed Spofforth's cause, despite her inability to verify her claims or explain the origin of the "Ali Al Shakati" name. This support appeared to stem from her connections to online right-wing and conspiracy-focused communities.

25. The engagements of Spofforth's tweets also raise questions about the potential and presence of bots and other forms of platform manipulation. It is still relatively trivial for fake accounts to mass follow and retweet strategic disinformation. Despite Elon Musk's claims to have got rid of bots, this is demonstrably untrue. This lack of ability to control manipulation is a key problem of social media manipulation, as they allow bad actors to manipulate discourse relatively easily.

26. Though both Spofforth and Channel 3 Now News eventually removed their tweets following widespread controversy, the damage was done. The false information had already fuelled additional speculative disinformation linking the attacker to Muslim communities. Despite her role in spreading unsubstantiated claims, Spofforth's Twitter account remains active. Both Spofforth and Farhan Asif attempted to deflect responsibility by claiming they had merely copied the false information from other sources, with Asif specifically pointing to a UK-based X account as his source.

27. Spofforth, despite shifting stories about where she obtained/made up the name Ali Al Shakati, defended her action by saying she stated, 'if this is true'. Again, this points to the difficulty of determining genuine or disingenuous behaviour when people are spreading false information. It is also worth noting that police arrested Farhan and Spofforth but took no recourse against other accounts, even though numerous accounts tweeted false information about Shakati.

28. Objectively, there is no reason not to take action against accounts selectively. Many accounts sharing the name were anonymous, pointing to the problem of a lack of accountability due to the anonymity function of many platforms including X. While anonymity may in theory be useful in some contexts, it is also extremely helpful for bad actors spreading disinformation. The implicit message here is that influential figures spreading dangerous information should remain anonymous to avoid police scrutiny (If Spofforth and Asif were singled out is there an implicit assumption that the greater the influence, the greater the culpability? In other words, the greater the reach (following), the greater the accountability.)

Elon Musk, EuropeInvasionn and Others

29. Platform leadership played a direct and documented role in amplifying harmful narratives. During the UK riots, Elon Musk played a significant role in amplifying controversial content and disinformation on X (formerly Twitter). With 194 million followers—making him the platform's most-followed account—his engagement had an outsized impact on how information spread. The data shows he posted about UK politics 46 times during this period, generating over 808 million impressions. Many of these posts drew inflammatory comparisons between Britain and authoritarian regimes.

30. Musk's interventions demonstrably shaped online discourse. In one notable instance, he responded to user @stclairashley's post about the Liverpool riots—content that had originated from Tommy Robinson and was shared through right-wing channels—with the provocative statement "Civil war is inevitable." The data shows his reply triggered a dramatic 540% increase in retweets of the original post within an hour.[15]

31. His influence extended to amplifying specific narratives. When he engaged with the hashtag #TwoTierKeir on August 6, which falsely claimed UK law enforcement was discriminating against white far-right protesters, he gave new life to what had been a declining trend. His involvement effectively functioned as a "personal algorithm booster," pushing fringe content into mainstream conversation.[16]

32. The impact of his platform management decisions compounded these issues. Since acquiring X, Musk has significantly reduced content moderation capabilities. Independent fact-checkers identified at least 50 of his own posts as containing misinformation, which collectively received more than 1.2 billion views without being flagged or removed. This included sharing a fabricated Telegraph article about UK "detainment camps" for rioters. His engagement with xenophobic content and conspiracy narratives helped legitimize false information during a politically sensitive period.[17] He also brought Tommy Robinson back to the platform, and Robinson was highly active in spreading disinformation in the Southport Riots. [18] This also raises the question of a concatenation of disinfluence - influential disinfluencers creating amplification chains that dominate the algorithm.

33. Another account that was key in the spread of disinformation during the Southport riots was @Europeinvasionn. Its most common themes are posts about

---

[15] https://www.ft.com/content/1843b68e-64cf-479f-b354-a7081257d42e
[16] https://www.tortoisemedia.com/2024/08/08/how-hashtag-twotierkeir-took-over-musks-x
[17] https://www.nbcnews.com/tech/misinformation/elon-musk-misleading-election-claims-x-views-report-rcna165599
[18] https://www.theguardian.com/technology/2023/nov/06/x-elon-musk-reinstating-katie-hopkins-tommy-robinson

immigrants and Muslims.[19]

34. Evidence suggests that @EuropeInvasionn is engaged in a coordinated influence operation. The account underwent a radical identity shift—changing its handle and erasing prior tweets—a common hallmark of deceptive social media campaigns. It was initially a cryptocurrency account, but scrubbed its tweets, and changed into a disinformation hub called Europe Invasion. It subsequently changed to @EuropeInvasions and later @UpdateNews724.These multiple changes and maintenance of absurdly high levels of engagement are suspicious.[20]

35. Legitimate accounts rarely engage in such drastic transformations, suggesting a deliberate attempt to repurpose an established digital presence for disinformation. Europe Invasion also promoted an account called Algorithm Coach - which claimed to show users how to game algorithms to promote content. This suggests a potential ability to game algorithms.

36. Further investigations have revealed tenuous links between @EuropeInvasionn and influencers based in the United Arab Emirates (UAE) and Turkey. Following public exposure, the account's operator claimed to be an individual named "Stefan K" from Montenegro. However, this identity remains unverifiable, and the account's true origins and affiliations remain uncertain.[21]

37. Data analysis covering the period from July 27 to August 7 demonstrates the extent of the account's influence. In ten days following the Southport killings, it spread multiple false stories about immigrants and Muslims, racking up millions of impressions.[22]

38. The impact of @EuropeInvasionn's disinformation became particularly evident during the Southport stabbing incident. Its false tweet about the Southport killer being Muslim was the most shared post within the first nine hours. The speed and reach of this tweet suggest premeditated disinformation tactics aimed at shaping public discourse early in the news cycle. The early presence of false narratives reinforces the role of the account as an active disinformation amplifier.

39. @EuropeInvasionn's content is centred on anti-Muslim and anti-immigrant rhetoric. Approximately 66% of the total 101,964,085 impressions generated by the account between the 27th of July and the 9th of August were tied to disinformation or hateful narratives. This indicates a strategic focus on polarizing topics designed to incite hostility and reinforce existing prejudices.

---

[19] https://www.rferl.org/a/europe-invasion-x-account-disinformation-xenophobia-immigration-x-account/33239067.html
[20] https://www.rferl.org/a/europe-invasion-x-account-disinformation-xenophobia-immigration-x-account/33239067.html
[21] https://www.rferl.org/a/europe-invasion-x-account-disinformation-xenophobia-immigration-x-account/33239067.html
[22] https://marcowenjones.substack.com/p/who-is-europe-invasionn-and-what

40. Despite the account's questionable origins and history of spreading falsehoods, it has received amplification from Elon Musk on multiple occasions, particularly during periods of civil unrest. Additionally, @EuropeInvasionn benefits from Twitter's "Blue Check" system, and therefore its falsehoods are algorithmically amplified.

41. Previously, verification required identity-based vetting to confirm the authenticity of notable figures, such as journalists and public officials. However, under the new Twitter Blue subscription model, verification can be purchased without identity verification, relying only on credit card authentication. This shift has created a "pay-for-play" system where anonymous or bad-faith actors can buy visibility and algorithmic amplification, allowing their content to spread more effectively.

42. The structural flaws in this verification model have been widely criticized. The European Union (EU) has initiated legal action against X (formerly Twitter) under the Digital Services Act (DSA).[23]

43. The Southport riots revealed the deployment of AI-generated content in disinformation campaigns. @EuropeInvasionn also utilized AI tools to create provocative but entirely fabricated images that enhanced the spread of xenophobic narratives. The combination of AI-generated content with algorithmic amplification creates particularly concerning dynamics. Bad actors can rapidly produce compelling false narratives that platform systems then promote based on engagement metrics. This represents a step-change in disinformation capabilities that current regulatory frameworks are not equipped to address.

44. Detailed analysis of impression data reveals critical timing patterns that demand regulatory attention. The most significant acceleration of harmful content occurred on July 29th, when false claims achieved viral spread before any fact-checking could occur. Initial false tweets from smaller accounts were rapidly amplified by larger "verified" accounts, creating a cascade effect that overwhelmed traditional fact-checking mechanisms.

UK response

45. Current UK regulatory mechanisms proved fundamentally inadequate in addressing the speed and scale of modern disinformation campaigns. When users reported content that explicitly called for violence against minorities and migrant housing, platform systems repeatedly failed to identify clear violations of UK law. This systematic failure highlights critical gaps between existing regulations and platform operations.

46. The case of Channel3NowNews exemplifies the international challenges facing regulators. This pseudo-news organization, eventually traced to Pakistan, created

[23] https://www.npr.org/2024/07/12/g-s1-9944/eu-takes-elon-musks-x-to-court-over-blue-checks-and-ads

professional-looking branding and systematically monetized disinformation. Despite operating from overseas, it significantly influenced UK public discourse, demonstrating how national regulations struggled to address transnational disinformation networks - who appear to know how to game algorithms for impressions (likely outrage).

47. The UK's response to social media harms requires significant strengthening of its regulatory framework, particularly in how Ofcom implements the Online Safety Act. A crucial first step is expanding platform oversight beyond simple size metrics to include functionality-based risk assessment. This would ensure that smaller but high-risk platforms like Telegram, which played a significant role in mobilizing offline violence during the Southport riots, are subject to appropriate duties and transparency reporting. Ofcom should establish mandatory baseline expectations and consistent measures to enable meaningful cross-industry comparisons of platform safety efforts.

48. The current regulatory framework needs stronger provisions for algorithmic accountability and data transparency. Independent third-party auditors should be granted access to platform data and policies to assess algorithmic ranking systems, while platforms must enhance enforcement of existing policies around children's safety and illegal content removal. The UK should align with the EU's Digital Services Act Article 40(12) through the proposed Data (Use and Access) Bill, mandating minimum data access requirements for researchers. This would address the current deterioration in data access, exemplified by the closure of Meta's CrowdTangle tool and prohibitive API costs on X, which have hampered systematic understanding of online harms.

49. Crisis response mechanisms require particular attention, as demonstrated by the rapid spread of disinformation during the Southport riots. The UK should implement explicit crisis protocols for platforms, similar to those in the EU's Digital Services Act, with enhanced monitoring and action requirements during critical events. These protocols must balance swift response capabilities with proper procedural accountability and human rights safeguards, ensuring platforms can effectively counter dangerous misinformation while maintaining appropriate oversight.

50. The emergence of AI-generated content presents new regulatory challenges that require immediate attention. The UK should develop clear guidelines for labelling AI-generated content and require platforms to implement robust measures to distinguish AI-generated political content from authentic material. This should be accompanied by increased transparency in the development of recommender algorithms and content moderation LLMs, supported by dedicated research investment in detecting LLM-generated content and identifying automated disinformation networks.

51. Finally, content moderation approaches need fundamental reform, shifting focus from reactive measures to proactive safety-by-design efforts. Platforms should be required to regularly review their practices and monitor policy enforcement

effectiveness, with mandatory training for moderation staff on emerging trends and technologies. This approach recognizes that while content moderation remains necessary, it cannot be the primary solution to online harms. The emphasis must be on preventing harmful content from gaining traction through platform systems in the first place, rather than attempting to remove it after it has already caused damage.

52. Failure to comply with such regulations needs to have substantial teeth, in the form of fines or other measures. As private companies, social media companies must be held accountable, especially when they claim to act in the public interest (e.g. free speech). This is distinctly different from a free market of speech, in which money, influence, and outrage, ensure what voices get heard.

26 February 2025