Platform Regulation and Hate Speech on YouTube

Çiğdem Işbuğa

Introduction:

The governance of online speech has increasingly shifted from public regulatory institutions towards privately owned digital platforms. Among this, Alphabet's YouTube has emerged as a key influence within the digital landscape. With over 2.70 billion monthly active users (Limelightdigital, 2025), YouTube holds a position of considerable influence through visibility of content and monetisation. As a result, research surrounding the regulation of harmful or extremist online speech are focused on the internal governance mechanisms developed by the company.

This report primarily aims to examine whether a platform structurally incentivised towards engagement can address the harms that said engagement produces. YouTube's model of self-regulation acts more as a form of private governance organisation where algorithmic transparency and commercial incentives inherently impose limitations on the effectiveness of hate speech content moderation.

Gillespie (2018) characterises content moderation as a central commodity offered by social media platforms, indicating that moderation systems often act as a mechanism to maintain service rather than function as neutral intermediaries. Platforms instead curate online speech through community guidelines. For example, YouTube's Hate Speech Policy currently states that it does not allow content that 'promotes violence or hatred against individuals or groups' based on protected characteristics classed under its policy (YouTube, 2026). Protective characteristics are set and defined by the company itself. These mechanisms position YouTube as a direct host and arbiter of online discourse contained on its platform. Therefore, the company determines which form of content remains visible or economically viable within its platform.

The neutrality of these governance systems has raised concerns among scholars, particularly surrounding oversight of hate speech content moderation practices. Diaz (2024) notes that 'colourblind policy' prevents action for the damage inflicted on marginalised communities. Hate speech policies only provide surface-level insight of who is a threat and it fails to account for the identification of contemporary white supremacist dog whistles and discourse. Accordingly, policies intended to appear neutral may inadvertently enable bigoted speech to persist within platforms.

In addition, economic incentives allude to another factor in moderation decision-making. Hokka (2020) indicates that YouTube's content creator revenue-sharing model results in restrictions only being implemented once reputational or advertiser pressure exceeds commercial benefits of the content (p. 155). While YouTube removes hate speech content, its moderation practices are shaped by commercial priorities and algorithmic systems. Such factors shape the visibility and regulation of hate speech on the platform.

This report is structured as follows; the first section positions the debates surrounding algorithmic amplification by interrogating whether YouTube's recommendation system produces political radicalisation or reflects pre-existing demand. The second section will analyse moderation as an economic process by examining the limits of the AI-driven moderation system. The final section addresses the overall limits of self-regulation before concluding on the implications of an unregulated platform.

Platform self-regulation and Algorithmic Recommendation

Debates surrounding YouTube's role in the circulation of extremist or hateful political content raises the prospect of whether the platform's algorithmic recommendation system actively pushes content or reflects pre-existing beliefs of users. This contrast is reflected in Munger and Phillips' (2022) description of the 'supply and demand' framework. The framework explains the spread of right-wing media on the platform, positioning the algorithm as a variable that mediates viewer interest in political content rather than solely generating interest. This analysis suggests that the majority of right-wing viewership on YouTube since 2019 is mainly attributable to the entry of mainstream conservative content. YouTube reflects this pre-existing demand rather than manufacturing interest itself.

Nonetheless, further studies on recommendation algorithms point to the processes of ideological radicalisation of users gradually being directed towards extreme content. Lewis' (2018) analysis of the Alternative Influence Network (AIN) presents a loosely connected ecosystem of political creators ranging from mainstream commentators, such as Ben Shapiro and Candence Owens to more explicitly ideological figures, such as Stefan Molyneux. Molyneux's content in particular incorporates white nationalist and eugenics messaging.

The AIN reflects how ideological content exists on a spectrum. Moreover, Lewis identifies that ideological testimonials, search engine optimisation, strategic controversy and political self-branding ultimately manoeuvre the recommendation system. This indicates that YouTube's recommendation system serves as a vital infrastructure that connects audiences to political adjacent creators through algorithmically mediated visibility.

Recent empirical research suggests that recommendation algorithms alone are enough to drive exposure to extremist content. Through the use of 100,000 sock puppet accounts, Haroon et al's audit-style study examined algorithmic influence isolated from user behaviour. Their findings indicate that concerns persist surrounding the personalisation measures aiming to maximise user engagement, resulting in the amplification of pre-existing political biases of users. In addition, ideologically congenial recommendations dominated especially for far-right users (p. 6). The study concludes that YouTube functions primarily as a medium allowing extremist content creators to reach an already radicalised community of users, reinforcing and rewarding while failing to resolve questions focused on addressing algorithmic responsibility.

Even if supplies of extremist content originated through the AIN, the recommendation architecture ultimately reinforces pre-existing ideology. This can influence users on either side of the political spectrum to easily shift to the extreme. Therefore, the algorithm functions less as a mechanism of radicalisation, but more as a system that underpins existing preferences to align with engagement and content consumption patterns.

The concept of whether moderation intervenes in instances of far-right radicalisation establishes a further dimension surrounding the platform's moderation governance. In 2019, YouTube introduced changes to its moderation system, aiming to crack down on violent extremism and supremacist content (YouTube, 2019). It announced that it was demoting content associated with conspiracy theories or extremist narratives in its algorithm. Subsequently, the platform banned prominent white supremacist channels in 2020, including David Duke, the former leader of the Klu Klux Klan (BBC, 2020).

The most prominent channel removal occurred prior to the announcement when Alex Jones' Infowars channel was banned in 2018. This adjustment in moderation focus does not necessarily indicate a reduction in ideological polarisation. Instead, it establishes the capacity of the platform's moderation practices. Studies on the effectiveness of deplatforming processes further reinforce this notion. Rauchfleisch and Kaiser (2024) analysed the trajectory of 11,000 far-right and conspiracy content creators across YouTube and video sharing platform, BitChute following the changes in YouTube policy approach. The analysis proposes that while 5.6% of the channels no longer exist on YouTube, the success of the content creators is directly connected to YouTube itself as a platform (p. 14).

Deplatforming limits reach in certain aspects, but it does not eliminate the underlying ideological content. This is particular clear during instances when such content is reuploaded to bypass bans and hate speech filters. Rauchfleisch and Kaiser underpin the migration of the content to alternative platforms within existing communities. High-profile removals, similar to the likes of banning of Alex Jones, demonstrates existing strengths and limitations within platform moderation. Such changes in policy restrict visibility while also generate attention thus encouraging audiences to either distribute the content themselves or move to a substitute platform.

Limitations in Moderation: Economic incentives and Automation

YouTube presents its moderation system as a neutral enforcement of community standards through its policies and guidelines. YouTube's most updated Hate Speech Policy states that content promoting violence or hatred against individuals based on protected group status under its policy is not permitted. Exceptions are made for content with 'educational, documentary, scientific or artistic context' (YouTube, 2026). The intention of the policy and community guidelines does not act as arbitration, rather serving more as a gesture (Gillespie, 2018). It serves as a gesture to users and more importantly, to lawmakers, assuring them that no further regulation is

necessary (p. 47). Despite YouTube presenting its moderation system and Hate Speech Policy as gestures of neutral enforcement of community standards, questions surrounding platform governance shaped by economic incentives persist.

Criticisms surrounding platform neutrality are mostly concerned with blind spots within moderation gestures. Diaz (2024) regards the 'myth of colorblind content moderation' as actively reproducing existing discriminatory practices and hierarchies. This is particularly clear through practices and policies treating white supremacy as an anomaly rather than a structural component. Nevertheless, YouTube's early tolerance of creators such as PewDiePie and Stefan Molyneux, illustrates this pattern, as creators whose content at various points included antisemitic material and eugenic concepts, remained on the platform. Internal company assessments in 2018 had even described PewDiePie's posts as driven by 'awkwardness' rather than hatred (Diaz, 2024).

It is important to note that at the time, PewDiePie was the most subscribed creator on YouTube, and such examples of 'awkward' posts included videos of him commissioning underpaid actors in India to hold signs with antisemitic phrases. (Hokka, 2020). This directly frames the policy blind spot as identified by Diaz: the inability to apply hate speech standards and simultaneously neglecting evolving white supremacist dog whistles. Bousier et al (2022) further extends this analysis to white supremacist historical revisionism on YouTube by documenting through the analysis of 27 videos totalling over three hours of content and identifying how creators systematically rewrite historical narratives from a racist perspective. More importantly, the study identifies how the platform's recommendation system directs users towards similar material.

Policy blind spots are not the only governance failure within YouTube's moderation practices. Monetisation and incentives situate governance failure within the platform's very own revenue-sharing model. As long as a creator's audience scale and advertiser relations remain commercially viable, the platform's commercial incentives will hold the potential to operate against decisive moderation. Hokka (2020) observes that YouTube being not interested in restricting the free speech of creators, regardless of their ideological message being harmful to society.

Moderation decisions are ultimately filtered through a hierarchy of commercial priorities: advertiser pressure, reputational management, and growth incentives. As a result, the consequence is a system in which moderation is far from being principled, responding to reputational scandals rather than structural harms to wider society. Barrett and Hendrix (2022) further document how YouTube's advertising revenue programme enables popular creators to generate income while amplifying hate speech and misinformation. Far-right content creators, organised misogynists, and supporters of oppressive regimes exploit the platform's reach to spread harmful content despite policy changes.

Moreover, the deployment of AI in content moderation introduces further structural moderation limitations. The introduction of automated moderation practices stem from the harms to human moderators and scale of content on the platform to get through (Gillsipie, 2018). Mattheis and Kingdon (2023) point to a range of tactics used by extremist creators to strategically adapt to moderation systems and exploit the boundaries of prohibited content. Machine learning models

trained to identify content from one extremist organisation may fail to detect material from another due to factors such as linguistic differences in propaganda.

Content creators who deliberately navigate the fine line between violating community guidelines and not crossing stated policy thresholds are able to maintain a presence while signalling ideological alignment to audiences. This adaptive dynamic indicates that AI moderation is outpaced by strategic actors. Conversely, an additional limitation highlights the wrongful removal of important content. The Censorship Effect (2022) documents how YouTube's automated systems have removed footage that is central to the prosecution of war crimes, rendering the content inaccessible to researchers and investigators. The analysis across both instances is consistent as moderation is reactive and enforcement is uneven. Therefore, the irregularity in consistence between YouTube's commercial responsiveness and regulatory implementation forms a gap in automated moderation practices.

Limits of Self-Regulation

Drawing upon the preceding analysis, several limitations of YouTube's self-regulatory model have been established. First, the 2019 algorithm changes did not resolve underlying dynamics. This is evident in Munger and Phillips' (2022) observation of how most of the right-wing content production and viewership since 2019 was driven by the entry of mainstream conservative creators, indicating that algorithmic demotion shifted the political composition of the platform without eliminating influence. The supply-and-demand dynamics that enable the Alternative Influence Network have demonstrated this. Second, deplatforming operates conditionally, as Rauchfleish and Kaiser (2024) exemplify its effectiveness being dependent on platform infrastructure. Third, automated moderation is powerless in the face of strategic adaptation by content creators who treat community guidelines as easily avoided constraints rather than enforced rules (Mattheis and Kingdom, 2023). Last, and most importantly, YouTube's governance remains commercially aligned, internal and opaque; there is no independent oversight mechanism, transparent process or structural separation between commercial interests and moderation decisions shaping content consumption.

Recent reports indicate that YouTube has quietly made further adjustments to its moderation practices in 2024 without any announcements (BU Today, 2025). The changes were partly linked to upping its threshold for how much potentially prohibited content a video can contain with a condition that the video is in public interest. Previously, videos could only contain a quarter of questionable content before removal. Such changes are better understood as YouTube changing moderation practices to fit the current mainstream conservative political landscape.

Another quiet change is linked to YouTube's new pilot programme in which previously terminated creators are handed an opportunity to rejoin the platform (The Express Tribune, 2025). One prominent creator benefiting from this programme is controversial live streamer Sneako, a previously banned creator in 2022 for repeatedly sharing content that spread misinformation about the COVID-19 pandemic and U.S. elections. YouTube's decision to reinstate previously

suspended channels directly corresponds to the inherent limitations of platform self-regulation, particularly where the absence of external accountability mechanisms allow moderation standards to be flexible over time. The rise of mainstream conservative content further contextualises this shift, as YouTube ultimately recalibrates its moderation practice in response to evolving patterns of engagement and demand. This is evident in cases such as Sneako, whose association with manosphere creators and continuation of controversial content raises question about the criteria in the pilot programme underpinning reinstatement decisions.

Beyond moderation and individual enforcement decisions, further limitations of platform self-regulation lie in the design of recommendation systems. Baker, Ging and Andreasen's (2024) study into the role of algorithmic functions on YouTube short in promoting male supremacist influencers involved the use of 10 sock puppet accounts on 10 blank smartphones. Findings from the study revealed that YouTube Shorts had a significantly high recommendation rate for toxic masculinity content. Accounts made specifically to mimic 16-year-olds and 18-year-olds were recommended this form of content after 17 and 2 minutes of viewing time, respectively.

The study also highlighted that 5.2% of recommend content on YouTube Shorts were reactionary right-wing and conspiracy content. This reinforces the contradictory nature of YouTube's governance model, as the platform simultaneously seeks to regulate harmful content while algorithmically incentivising its visibility. With no external mechanism to resolve tensions between commercial interest and social responsibility, YouTube's moderation practices reflect the disparity between its commitments to platform safety and its economic dependence on the visibility of controversial content.

Conclusion

This report has examined YouTube's model of platform self-regulation, identifying algorithmic opacity, selective enforcement and commercial incentives as limiting the effectiveness of hate speech moderation. Evidence drawn from algorithmic audit research (Haroon et al, 2023), deplatforming analysis (Rauchfleisch and Kaiser, 2024), enforcement critical analysis (Diaz, 2024; Hokka, 2020), and examinations of AI moderation limits (Mattheis and Kingdom, 2023; Barrett and Hendrix, 2022), demonstrate that YouTube's regulatory limitations are not incidental and instead structural in nature. This implications of this analysis extend beyond questions of platform governance into concerns about private power and accountability.

Additionally, the absence of independent oversight and commercial alignment of moderation decisions raises grounded questions about whether YouTube's practices are sufficient. The platform quietly rolling back on moderation practices and reinstating suspended creators further fuels such concerns. The evidence examined in this report suggests that the conditions for self-regulation and moderation are not yet in place in its current form.

References:

Barrett, P.M. and Hendrix, J. (2022), 'A platform 'weaponized': How YouTube spreads harmful content–and what can be done about it'. *NYU Stern School of Business Research Paper Forthcoming*, doi: https://dx.doi.org/10.2139/ssrn.5495142

BBC. (2020), 'YouTube bans prominent white supremacist channels', Available at: https://www.bbc.co.uk/news/business-53230986 , [Accessed: 20/02/2026],

Boursier, T., Kaiser, C., Khiari, O. and Lühr, V.S. (2022), 'White Supremacism on YouTube: How to Rewrite History from a Racist Point of View', *Temporalities of Diversity/Temporalités de la diversité/Zeitlichkeiten der Vielfalt*, pp.65-87. Available at: https://books.google.co.uk/books?hl=en&lr=&id=spOWEAAAQBAJ&oi=fnd&pg=PA65&dq=Bousier+youtube&ots=geg9EmTKnb&sig=X_7WpTR2ZPY4v1u13SUWKqLyRdo&redir_esc=y#v=onepage&q&f=false

Diaz, A. (2023), 'Online racialization and the myth of colorblind content policy', *Boston University Law Review Rev*, 103:1929, Available at: www.bu.edu/bulawreview/files/2023/12/DIAZ.pdf

Ging, D., Baker, C. and Brandt Andreasen, M. (2024), *Recommending Toxicity: The role of algorithmic recommender functions on YouTube Shorts and TikTok in promoting male supremacist influencers*. DCU Anti-Bullying Centre: Dublin, available at: https://doras.dcu.ie/31681/

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press.

Haroon, M., Wojcieszak, M., Chhabra, A., Liu, X., Mohapatra, P. and Shafiq, Z. (2023), 'Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations', *Proceedings of the national academy of sciences*, 120(50), doi: https://doi.org/10.1073/pnas.2213020120

Hokka, J., (2021), 'PewDiePie, racism and Youtube's neoliberalist interpretation of freedom of speech', *Convergence*, 27(1), pp.142-160, doi: https://doi.org/10.1177/1354856520938602

Lewis, R. (2018). 'Alternative influence: Broadcasting the reactionary right on YouTube', *Data & Society Research Institute*, Available at: https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf

LimeLightDigital, (2025). 'How Many People Use YouTube? (2026 User Statistics)', Available at: https://www.limelightdigital.co.uk/youtube-statistics/ , [Accessed: 20/02/2026]

Munger, K., & Phillips, J. (2022), 'Right-Wing YouTube: A Supply and Demand Perspective', The International Journal of Press/Politics, 27(1), p. 186-219, doi: https://doi.org/10.1177/1940161220964767

Ottman, B., Davis, D., Ottman, J., Morton, J., Lane, J.E. and Shults, F.L., (2022). 'The Censorship Effect: An analysis of the consequences of social media censorship and a proposal for an alternative moderation model', *Minds.com,* Available at: https://cdn-assets.minds.com/The_Censorship_Effect.pdf

Rauchfleisch, A. and Kaiser, J., (2024), 'The impact of deplatforming the far right: an analysis of YouTube and BitChute'. *Information, Communication & Society*, *27*(7), pp.1478-1496, doi: https://doi.org/10.1080/1369118X.2024.2346524

YouTube. (2019), 'Our ongoing work to tackle hate', Available at: https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate/ , [Accessed: 20/02/2026],

YouTube. (2026) 'Hate speech policy', Available at: https://support.google.com/youtube/answer/2801939?hl=en-GB , [Accessed: 20/02/2026],

ES ETHICALSCREENING

www.ethicalscreening.co.uk

01242 539 850

info@ethicalscreening.co.uk

www.linkedin.com/company/ethical-screening-limited

Ethical Screening is the trading name of Ethical & Environmental Screening Services Ltd.
Directors: Michael Head and Gerard Llewellyn
Registered Office: Formal House, 60 St. George's Place, Cheltenham, GL50 3PN
Registered in England & Wales.
Registration number: 3633308
VAT Registration number 713760544