How Algorithms Influence Harmful Online Conduct

Submitted by the Anti-Defamation League to the Joint Committee on the Draft Online Safety Bill
September 15, 2021

I. About the Anti-Defamation League (ADL)

Since 1913, ADL's mission has been to "stop the defamation of the Jewish people and to secure justice and fair treatment for all." Dedicated to combating antisemitism, prejudice, and bigotry of all kinds, as well as defending democratic ideals and promoting civil rights, ADL is a leading voice in fighting hate in all forms, including online. ADL has gained particular experience in this space since we launched our Center for Technology and Society (CTS) in 2017. CTS leads the global fight against online hate and harassment. In a world riddled with antisemitism, bigotry, extremism, and disinformation, CTS acts as a fierce advocate for making digital spaces safe, respectful, and equitable for all people.

II. Summary

ADL's submission of evidence answers the question posed by the Joint Committee on the Draft Online Safety Bill's call for evidence: What role do algorithms currently play in influencing the presence of certain types of content online and how it is disseminated? This submission will cover three primary topics: (1) the role algorithms play to fuel Big Tech's business model; (2) the role algorithms play in spreading hateful and harassing content and the impact on targets; (3) the role algorithms play in amplifying misinformation and spreading extremism, which can lead to physical violence. Finally, the submission will offer recommendations considering the information discussed. Much of our research related to this topic focuses on the American experience of online hate and extremism. Here, we provide it with the hope that it can be instructive in assessing similar issues in the United Kingdom.

III. Evidence

A. Platforms' algorithms spread harmful information because it is good for their business model

1. Artificial intelligence (AI) and algorithms play a powerful role in the dissemination of online harm. "AI can be understood as machines that predict, automate, and optimize tasks in a manner that mimics human intelligence, while [machine learning] algorithms, a subset of AI, use statistics to identify patterns in data."[1] Social media platforms use algorithms, largely fueled by AI and machine learning (ML) systems, to deliver and moderate content, to determine what content should be recommended to a user, and to serve advertisements to users.

[1]"Trained for Deception: How Artificial Intelligence Fuels Online," Coalition to Fight Digital Deception, last modified September 14, 2021, https://static1.squarespace.com/static/6103ea02f6f50e4407fa34cf/t/613fd03e1b58a82cf447445c/1631572030813/Trained_for_Deception_How_Artificial_Intelligence_Fuels_Online_Disinformation.pdf.

Algorithms make these highly personalized decisions by collecting and synthesizing vast amounts of user data.

2. One primary reason algorithms contribute to and influence the presence of harmful online content on social media is that platforms are founded on a business model that optimizes for user engagement.[2] When a user interacts with a piece of content, algorithmic systems recognize signals, like popularity, and then amplify that content. If content is forwarded, commented on, or replied to, social media algorithms almost immediately show such content to more users, prompting increased user engagement, and thus increasing advertising revenue. Research shows that controversial, hateful, and polarizing information and misinformation are often more engaging than other types of content and, therefore, receive wider circulation.[3]
3. A 2018 MIT study showed false news spreads more rapidly than the truth on social media. "Falsehood diffused significantly farther, faster, deeper, and more broadly than the truth in all categories of information, and the effects were more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information."[4] Another study, published in *Nature* concluded, "Based on engagement, Facebook's Feed drives clicks and views, but also privileges incendiary content, setting up a stimulus–response loop where outrage expression becomes easier and even normalized."[5]
4. Hate speech, conspiracy theories and misinformation, powered by algorithms, amplify divisive and false content through news feeds. Algorithms feed users tailored content, based on factors including browsing activity. If a user has viewed or searches for hateful content, algorithms learn to serve the same user similar or more extreme content. Big Tech's fundamental business model exploits people's predilection for clicking on incendiary content and sharing misinformation and divisive material.

B. **Algorithms contribute to the spread of online hate and harassment**

5. Two primary ways social media platforms' algorithms contribute to the spread of hateful and harassing content are (1) platforms' overreliance on algorithmic AI/ML systems to detect and moderate hateful content and (2) the fact that once hateful and harassing content evades detection from content moderation systems, it is then spread and amplified by platforms' ranking and recommendation algorithms (because the content has such high engagement rates). Because hate and harassment disproportionately target marginalized groups, the persistent

[2] "The Simplest Way to Rein In Facebook and Big Tech," Jesse Lehrich, last modified March 29, 2021, https://crooked.com/articles/facebook-big-tech-surveillance/.
[3] "Facebook's Hate Speech Problem Is Even Bigger Than We Thought," Anti-Defamation League, last modified December 23, 2020, https://www.adl.org/blog/facebooks-hate-speech-problem-is-even-bigger-than-we-thought.
[4] "The Spread of True and False News Online," Science, last modified March 9, 2018, https://www.science.org/doi/abs/10.1126/science.aap9559.
[5] "Angry by Design: Toxic Communication and Technical Architectures," Humanities and Social Sciences Communications, last modified July 30, 2020, https://www.nature.com/articles/s41599-020-00550-7.

presence of harassment, bigotry and conspiracy theories on social media platforms has deeply impacted vulnerable and marginalized communities including people of color, religious minorities, the LGBTQ+ community and others.[6]

6. Most major social media companies have policies that determine the types of content, user accounts, and groups permitted on their platforms. To enforce these policies, sometimes called "Community Guidelines" or "Terms of Service," most platforms use both human and AI/ML tools to moderate content.[7] "While overbroad content moderation raises freedom of expression concerns, content moderation is important for addressing misinformation, disinformation, harassment, and racist or hateful content online."[8]
7. "AI can improve the effectiveness of human moderators by prioritizing content to be reviewed by them based on the level of harmfulness perceived in the content or the level of uncertainty from an automated moderation stage."[9] It is important to note, however, that an overreliance on algorithmic content moderation will leave harmful and harassing content proliferating on platforms. Additionally, AI/ML systems often reproduce existing social biases because they are fueled by biased data and assumptions.[10] A key limitation of automated content moderation is that "the difficulty of detecting harmful content is that the intention of the content is often nuanced."[11] Specifically, it's often the *context* rather than the content that causes harm to others.[12] Notably, white nationalist groups are particularly skilled at developing new terms or memes in order to avoid automated content moderation.[13]
8. Based on reports from 2020, Facebook's "algorithms and policies did not make a distinction between groups that were more likely to be targets of hate speech

---

[6] "Online Hate and Harassment: The American Experience 2021," Anti-Defamation League, last modified March 2021, https://www.adl.org/online-hate-2021; "The Trolls are Organized and Everyone's a Target," Anti-Defamation League, last modified October 2019, https://www.adl.org/trollsharassment.

[7] "ADL Calls for Platforms to Take Action to Address Hate Online During Pandemic," Anti-Defamation League, last modified May 8, 2020, https://www.adl.org/blog/adl-calls-for-platforms-to-take-action-to-address-hate-online-during-pandemic.

[8] "Trained for Deception," https://static1.squarespace.com/static/6103ea02f6f50e4407fa34cf/t/613fd03e1b58a82cf447445c/1631572030813/Trained_for_Deception_How_Artificial_Intelligence_Fuels_Online_Disinformation.pdf.

[9] "Use of AI In Online Content Moderation," Cambridge Consultants, last modified 2019, https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf.

[10] Safiya Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, (New York: NYU Press, 2018).

[11] "How AI Can Help to Moderate Content," Forbes, last modified December 1, 2020, https://www.forbes.com/sites/junwu1/2020/12/01/how-ai-can-help-to-moderate-content/?sh=2a3077625c2b.

[12] "Content or Context Moderation? Artisanal, Community- Reliant and Industrial Approaches," Data and Society, last modified November 14, 2018, https://datasociety.net/library/content-or-context-moderation/.

[13] "I Testified at a Congressional Hearing on White Nationalism. Here's Some of What I Wish We Had Discussed," Anti-Defamation League, last modified April 25, 2019, https://www.adl.org/blog/i-testified-at-a-congressional-hearing-on-white-nationalism-heres-some-of-what-i-wish-we-had.

versus those that have not been historically marginalized."[14] ADL's series of social media platform "report cards" also illustrate that platforms need to strengthen their content moderation and enforcement systems.[15] Platforms can and must reduce bias and improve the effectiveness of content moderation systems by increasing resources focused on policy and product improvements from a civil and human rights lens. The Stop Hate for Profit Campaign recommended that platforms "establish and empower permanent civil rights infrastructure including C-suite level executive with civil rights expertise to evaluate products and policies for discrimination, bias, and hate."[16]

9. Additionally, platforms must invest in more human content moderation to work in tandem with AI systems. In the wake of the coronavirus pandemic, many tech companies sent home human content moderation teams, leaving platforms to be predominantly moderated by automated systems—and often automated systems are tuned towards leniency.[17] Increased reliance on automated systems has not proven to be effective at moderating content that requires context, including a significant portion of antisemitic content. For example, in April 2020, New York City Mayor Bill de Blasio tweeted about the Jewish community, and many clearly antisemitic tweets followed in reaction and remained active on Twitter for extended periods of time, despite having been reported to the platform.[18] "There are numerous examples of internet platforms relying on biased AI and ML-based tools to make content moderation decisions that have resulted in harmful and discriminatory outcomes."[19]
10. ADL's 2021 Online Hate and Harassment Survey, which surveyed a nationally representative group of Americans, revealed some of the inflammatory content perpetuated on social media.[20] According to ADL's latest data, 41% of Americans have experienced online harassment, and 1 in 3 of those harassed attributed at least some harassment to a protected, identity characteristic. And this hate isn't

[14] "Facebook to Start Policing Anti-Black Hate Speech More Aggressively Than Anti-White Comments, Document Shows," The Washington Post, last modified December 3, 2020, https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/.

[15] "2021 Online Antisemitism Report Card," Anti-Defamation League, last modified 2021, https://www.adl.org/resources/reports/2021-online-antisemitism-report-card; "Online Holocaust Denial Report Card: An Investigation of Online Platforms' Policies and Enforcement," Anti-Defamation League, last modified 2021, https://www.adl.org/holocaust-denial-report-card.

[16] "Recommended Next Steps," Stop Hate for Profit, last modified January 2021, https://www.stophateforprofit.org/productrecommendations.

[17] "ADL Calls for Platforms to Take Action to Address Hate Online During Pandemic," Anti-Defamation League, last modified May 8, 2020, https://www.adl.org/blog/adl-calls-for-platforms-to-take-action-to-address-hate-online-during-pandemic.

[18] Id.

[19] "Trained for Deception,"https://static1.squarespace.com/static/6103ea02f6f50e4407fa34cf/t/613fd03e1b58a82cf447445c/1631572030813/Trained_for_Deception_How_Artificial_Intelligence_Fuels_Online_Disinformation.pdf.

[20] "Online Hate and Harassment: The American Experience 2021," https://www.adl.org/online-hate-2021.

merely taking place on fringe message boards. 75% of those harassed said at least some harassment happened on Facebook. 24% reported harassment on Twitter, 21% on YouTube, and 15% on Snapchat.

11. Importantly, 27% of the respondents surveyed experienced severe online harassment, which includes sexual harassment, stalking, physical threats, swatting, doxing, and sustained harassment. Of this 27%, 52% of the respondents identified as LGBTQ+, 36% identified as Muslim, 29% identified as female, 25% identified as male, 23% identified as African American, 22% identified as Jewish, 21% identified as Hispanic or Latino, and 17% identified as Asian-American.
12. Identity-based harassment is worrisome and affects the ability of already marginalized communities to be safe in digital spaces when this harassment goes undetected, or worse, is amplified by a platform's own algorithms. ADL's work with victims of online hate and harassment revealed to us that such experiences have a chilling effect on a victim's participation in civic engagement.[21] Speech rights exist on all sides of this equation. Such rights are not just the concern of the loudest or most empowered speakers. On the contrary, our legitimate concerns over free speech need to take into consideration who is silenced or sidelined.[22]
13. Online hate directed at young people is also prevalent. ADL's latest survey, a nationally representative sample of the 97 million American individuals who play online multiplayer games, reveals this phenomenon.[23] The survey found that three out of five young people (ages 13-17) who play online games experienced harassment in online multiplayer games. Among the young gamers who were targeted based on their in-game appearance, 13% reported being targeted every time they played. Among those young gamers excluded from joining a game or chat based on their identity, 15% had this experience every time they played. 37% of female-identified young gamers experienced harassment as a result of their gender.
14. The harassment that young people experience in online multiplayer games affects their online and offline lives. According to the report, almost one in three young gamers who experienced harassment in online multiplayer games quit specific games.[24] A third of young gamers changed how they play, including not speaking in voice chat and altering their usernames. Voice chat is notorious for being a significant locus of in-game abuse. A username that refers to a player's gender, race, or other identity characteristic also can serve as a target for harassment. Finally, in-game harassment has offline consequences for young people. 16% of

[21] "The Trolls are Organized and Everyone's a Target," Anti-Defamation League, last modified October 2019, https://www.adl.org/trollsharassment.
[22] "I Testified at a Congressional Hearing on White Nationalism," https://www.adl.org/blog/i-testified-at-a-congressional-hearing-on-white-nationalism-heres-some-of-what-i-wish-we-had.
[23] "Hate is No Game: Harassment and Positive Social Experiences in Online Games 2021," Anti-Defamation League, last modified September 15,2021, https://www.adl.org/hateisnogame.
[24] Id.

young gamers in the U.S. reported that they treated people worse than usual after being harassed, and 10% reported their school performance declined.

C. **Algorithms spread misinformation and amplify extremism that can lead to on-the-ground violence**

15. Extremist groups are undoubtedly empowered by access to the online world and one key contributor is the existence of powerful algorithms that amplify the hateful voices of a few to reach millions around the world.[25] Major social media platforms, which employ algorithms designed to promote engagement and end up amplifying the most corrosive content, often recommend material that glorifies hate and violence.[26]
16. There is a clear connection between online antisemitic, racist, and hateful images and tropes reverberating on social media and offline hate and violence directed at marginalized communities.[27] In the United States, calls to violence in the name of white supremacy and "The Great Replacement," which has proliferated online and been amplified through algorithms, correlate to fatal shootings in Poway, El Paso, Pittsburgh and more, and led to the injuries and deaths at the white supremacist attacks in Charlottesville in 2017 and on the United States Capitol on Jan 6, 2021.[28] Further, the deadly insurrection at the United States Capitol is a key example of the violence that can erupt when extremist disinformation spreads on social media.
17. Fringe platforms, despite having relatively small userbases, leverage mainstream platforms like Twitter and Facebook to increase their reach and influence. Still, Big Tech platforms play a significant role in amplifying white supremacist content. Algorithms on Big Tech platforms operate with stunning speed, scope and impact. [29] Last fall a single "Stop the Steal" Facebook group gained more than 300,000 members within 24 hours, even after being reported. Thousands of new

---

[25] *ADL: Hearing on Hate Crimes and the Rise of White Nationalism, Before the House Committee on the Judiciary*, 116th Cong. (2019) (statement of Eileen Hershenov, Senior Vice President of ADL), https://docs.house.gov/meetings/JU/JU00/20190409/109266/HHRG-116-JU00-Wstate-HershenovE-20190409.pdf.

[26] *Examining the Domestic Terrorism in the Wake of the Attack on the U.S. Capitol, Before the House of Representatives Homeland Security Committee*, 117th Cong. (2021) (statement of Jonathan Greenblatt, CEO and National Director of ADL), https://www.adl.org/media/15833/download.

[27] "Moonshot & ADL Project Finds Anti-Black, Antisemitic, White Supremacist Internet Searches in Conjunction with Major Offline Events," Anti-Defamation League, last modified June 16, 2021, https://www.adl.org/news/press-releases/moonshot-adl-project-finds-anti-black-antisemitic-white-supremacist-internet.

[28] "Letter to Oversight Board," Anti-Defamation League, last modified June 2, 2021, https://www.adl.org/media/16436/download.

[29] *Domestic Terrorism and Violent Extremism: Examining the Threat of Racially, Ethnically, Religiously and Politically Motivated Attacks , Part II, Before the Senate Homeland Security and Government Affairs Committee*, 117th Cong. (2021) (Statement of Jonathan Greenblatt, CEO and National Director of ADL), https://www.hsgac.senate.gov/imo/media/doc/Testimony-Greenblatt-2021-08-05.pdf.

members a minute joined this group and some of them openly advocated civil war.[30]

18. Social media algorithms recommend content to extremist-leaning users, including related groups and pages that contain harmful content. According to an ADL report released in February 2021, data indicate that exposure to videos from extremist or white supremacist channels on YouTube remains disturbingly common. The study found that approximately one in ten participants viewed at least one video from an extremist channel (9.2%) and approximately two in ten (22.1%) viewed at least one video from an alternative channel, which were considered channels that can serve as gateways to more extreme forms of content.[31] Moreover, when participants watch these videos, they are more likely to see and follow recommendations to similar videos.
19. In recent years, extremists' online presence has reverberated across a range of social media platforms. This extremist content is intertwined with hate, racism, antisemitism, and misogyny—all of which are present in white supremacist ideology. Such content is enmeshed in conspiracy theories and spreads on platforms where algorithms are tuned to spread disinformation.[32]

IV. **Recommendations**:

ADL recommends several steps governments should take to combat the threat of online extremism fueled by algorithms. Government can pass laws and undertake other approaches that require regular reporting, increased transparency and independent audits regarding content moderation, algorithms, and engagement features. Additionally, legislators can make sure laws cover cybercrimes such as doxing, swatting, cyberstalking, cyberharassment, non-consensual distribution of intimate imagery, video-teleconferencing, and unlawful and deceptive synthetic media (sometimes called "deep fakes"). Governments could also create a publicly funded, independent, nonprofit centers that track online extremist threats in real-time and make referrals to social media companies and law enforcement agencies when appropriate. ADL's PROTECT[33] and

---

[30] "A GOP-Linked 'Stop the Steal' Facebook Group Is Gaining Thousands of Members A Minute," Vice, last modified November 5, 2020, https://www.vice.com/en/article/xgzx8q/a-gop-linked-stop-the-count-facebook-group-is-gaining-thousands-of-members-a-minute; "Extremists, Others Respond to President Trump's Calls to "Stop the Count,"" Anti-Defamation League, last modified November 6, 2020, https://www.adl.org/blog/extremists-others-respond-to-president-trumps-calls-to-stop-the-count.

[31] "Exposure to Alternative & Extremist Content on YouTube," Anti-Defamation League, last modified February 2021, https://www.adl.org/resources/reports/exposure-to-alternative-extremist-content-on-youtube#executive-summary

[32] *Domestic Terrorism and Violent Extremism: Examining the Threat of Racially, Ethnically, Religiously and Politically Motivated Attacks , Part II, Before the Senate Homeland Security and Government Affairs Committee*, 117th Cong. (2021) (Statement of Jonathan Greenblatt, CEO and National Director of ADL), https://www.hsgac.senate.gov/imo/media/doc/Testimony-Greenblatt-2021-08-05.pdf.

[33] "PROTECT Plan to Fight Domestic Terrorism," Anti-Defamation League, last modified 2021, https://www.adl.org/protectplan.

REPAIR[34] Plans note the role social media—and algorithms specifically—play in amplifying online harm and provide a framework to push hate and extremism to the fringes of the digital world. While these plans are U.S.-focused, ADL strongly believes these principles can be applied to decrease harmful online content in the United Kingdom and across the globe.

[34] "REPAIR Plan: Fighting Hate in the Digital World," Anti-Defamation League, last modified March 2021, https://www.adl.org/repairplan.