Written evidence submitted by Reset

Response to the DCMS Sub-Committee on Online Harms and Disinformation

September 2021

Introduction

1. **Reset, and many of our partner organisations across civil society, are advocates of the online safety agenda and have been for many years.** The public overwhelmingly support the end of self-regulation by tech platforms[1]. Other nations are watching closely, urging us to set a high bar. This legislation could not be more timely.

2. The draft Bill is not perfect. The language is vague in many places and raises more questions than answers. **It does not go as far as it should in protecting adults online, and leans too much towards *content* regulation rather than the *systems* approach of the White Paper.** Despite the lack of clarity, the intentions of the Bill are admirable. Many of its issues can be corrected through targeted revision.

3. To succeed, **the Bill must tackle the design features and algorithms which amplify harm.** If it becomes a takedown regime hinging on criminalising or deleting legal content, it will fail from both a practical perspective (it's impossible to delete our way out of the problem) and on freedom of expression grounds.

4. It is well documented, not least by the tech companies themselves, that **algorithms are the key driver of division and harm on online platforms**. As an internal Facebook presentation from 2018 stated: ***Our algorithms exploit the human brain's attraction to divisiveness***.[2] Recent research by Mozilla into YouTube's recommendation algorithms backs this up, concluding that videos which users regret watching "are primarily a result of the recommendation algorithm, meaning videos that YouTube chooses to amplify, rather than videos that people sought out".[3]

5. This predilection for the extreme serves the attention optimization business model of tech giants who want to secure maximum engagement in order to sell ads. The content which drives the most engagement is that which is provocative but not necessarily illegal. Mark Zuckerberg elaborates on this in a blog post, stating that:

[1] Online Nation – 2020 report, p37, Ofcom, 2020
[2] Facebook Executives Shut Down Efforts to Make the Site Less Divisive, Wall Street Journal, 26 March 2020
[3] YouTube Regrets, Mozilla Foundation, July 2021

"One of the biggest issues social networks face is that, when left unchecked, people will engage disproportionately with more sensationalist and provocative content. [...] At scale it can undermine the quality of public discourse and lead to polarization."[4]

The graph accompanying his blog post (Fig. 1) visualises this reality.

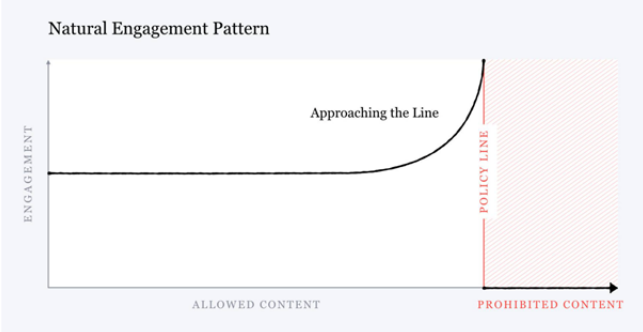


Fig. 1

6. Digital platforms tweak their algorithms all the time to drive engagement and develop new products. But, as Facebook's 'P (Bad for the world)' experiment shows, when changes to the algorithm *reduce* engagement, they are shelved - even when they reduce harm.[5] **The Bill must mandate the application of design features which reverse the curve in Fig.1, as well as ensuring platforms no longer have sole responsibility for setting their own 'Policy line'.**

Response to key questions

Does the draft Bill focus enough on the ways tech companies could be encouraged to consider safety and/or the risk of harm in platform design and the systems and processes that they put in place?

7. **In short, no.** The draft Bill acknowledges the power of design choices and algorithms in promoting harmful content. In the case of illegal content or content that is harmful to children (Clauses 9 and 10), services must operate "systems and processes" which mitigate against these risks, including against the dissemination of such content. For legal content (Clause 11), companies have no obligation to operate risk management via systems and processes and are free to manage such content how they choose as long as they set out their approach in their Terms and conditions ("T&Cs"). Platforms would be free to continue handling legal content how they see fit: deleting content en masse, leaving abuse and hate spiralling out of control, or doing nothing at all. This risks

[4] A blueprint for governance and enforcement, blog post, Mark Zuckerberg, 2018 (updated 2021)
[5] Facebook Struggles to Balance Civility and Growth, New York Times, 24 Nov 2020

perpetuating the status quo, codifying the power of corporations to determine what can be said online. It makes the Bill a *content* Bill which permits the deletion of legal content, rather than a *systems* Bill. And it also creates tiers of harms, with this category of content subject to the weakest harm reduction measures

8. **The Bill should be commended for including "content that is harmful to adults" or 'legal harms'.** This captures much of the abuse and hate witnessed by many on a daily basis such as COVID disinformation, bullying, climate change denial, pro-suicide and self-harm material, none of which is illegal and all of which can have a devastating impact. Just look at the racist abuse targeted at footballers after the European Cup final, much of which took the form of emojis.[6] While most agree that the hostility towards the black community is abhorrent, no-one wants to outlaw emojis. This presents a challenge to policymakers about how to tackle such harms without infringing on freedom of speech. The Bill can, and must, preserve freedom of speech while reducing online harms. It must not rely on removing or criminalising legal content but rather on service design choices which reduce the amplification and reach of harmful material. **Legal harms and adults' risk assessment duties must remain in the Bill, and must be subject to more rigorous harm reduction obligations.**

9. **Rather than asking companies to write rules for *content*, Clause 11 of the draft Bill should require them to improve their *systems and designs*: mandating practical solutions to minimise the spread of harmful material by focusing on preventative measures such as reduced amplification, demonetisation and targeting**. **This should be supported by a Code of Practice on harm reduction by design.** These measures introduce friction into an otherwise frictionless system. There are many examples of how this can be done well. This would mean that abusive tweets sent in the heat of the moment to a footballer who had a bad game aren't promoted to other disappointed fans, causing an abusive pile on. Before telling a Love Island heartthrob who has fallen from grace to kill themselves, users are asked to think twice. Having the option to delay when your comment is posted, becomes the norm. Being directed to authoritative, fact-checked sites about climate change or coronavirus before you watch a conspiracy theory video might give pause for thought. More examples of harm reduction by design in practice are at the end of this response.

10. **To support the enforcement of a harm reduction regime, OFCOM must be given clear powers to inspect the algorithms and systems promoting all forms of harmful content.** At present, the language in Chapter 5 of the draft Bill states that OFCOM can serve an "Information notice" to a service, requiring a person "to provide information which OFCOM believes that person has or is able to generate or obtain" (70). They may also "appoint a skilled person" (Clause 74) to help them with an investigation. There is no language in the Bill which excludes OFCOM from initiating audits of services' algorithms. However, recent commentary suggests OFCOM does not

[6] This Is Why They Took The Knee – HOPE not hate

view algorithmic audit as a power in its toolkit based on the draft Bill. This should be clarified, with **OFCOM being granted unequivocal authority to investigate algorithms as a driver of harm for all categories of content.**

11. Transparency powers in the Bill should also include a requirement for platforms to **share relevant data with accredited researchers studying online harms/safety**. This would give academia a much clearer picture of how harmful content is generated and promoted online. This in turn would help policymakers and the safety tech industry develop innovative ideas and products based on evidence and data.

12. **Transparency and oversight are the cornerstones of all existing regulatory regimes, be they in financial services, pharmaceuticals or the automotive industry. Ofcom must be given the appropriate powers to make the online safety regime sufficiently robust.**

What are the key omissions to the draft Bill, such as a general safety duty or powers to deal with urgent security threats, and (how) could they be practically included without compromising rights such as freedom of expression?

Disinformation

13. **Reset believes the Bill must meaningfully tackle disinformation.** The Government's commentary on the Bill suggests that COVID-19 disinformation will be in scope (perhaps via the powers granted to the Secretary of State in Clause 112) but that the intentions are not to include disinformation as a whole. **This creates a major risk to the UK, and undermines the ambition to make the UK the safest place to be online.**

14. The UK has witnessed or been subject to multiple coordinated disinformation campaigns in recent years. **Online disinformation campaigns which spill into offline harassment of journalists[7]; state backed disinformation campaigns inauthentically amplifying partisan views on Scottish referendum[8]; climate change denial[9]; disinformation discrediting the Security Services' investigations into the Sergei Skripal poisoning[10]; and 5G conspiracy theories[11] are just some of the attempts to undermine trust in authorities and sew confusion**.

15. The omission of disinformation from the draft Bill is at odds with original proposals in the Online Harms White Paper, as well as with the Home Office's recent consultation on Hostile State Activity which recognised the role of disinformation in undermining democracy. The DG of the Security Service reiterated this threat in his 2021 Annual Threat update.[12] At an international level, the recent communique following the G7

---

[7] Anti-vaxxers harass Jon Snow as they storm ITN headquarters, The Telegraph, 23 August 2021
[8] Facebook shuts fake Scottish independence accounts | Scotland, The Times, 6 March 2021
[9] Facebook's Climate of Deception: How Viral Misinformation Fuels the Climate Emergency, Avaaz
[10] Sergei Skripal and the Russian disinformation game, BBC News, September 2018
[11] Here's where those 5G and coronavirus conspiracy theories came from, FullFact

Summit committed G7 nations to "strengthening the G7 Rapid Response Mechanism to counter foreign threats to democracy including disinformation"[13]; and in The New Atlantic Charter, signed in June 2021, the UK and the US agreed that they "oppose interference through disinformation or other malign influences, including in elections"[14]. Meanwhile, the relevant areas of UK government policy, such as the Elections Bill, are silent on disinformation. **The Online Safety Bill must put these commitments into practice if it is to protect the UK against disinformation campaigns from domestic and foreign actors.**

Collective harm

16. To achieve this, the Bill must **include a definition of harm which accounts for the collective or societal impact of harmful content.** As COVID disinformation has highlighted, the impact of disinformation is absolutely collective in nature. As currently drafted, the Bill focuses on harm to the individual. This differs from the language in the Online Harms White Paper which proposed "prioritising regulatory action to tackle harms that have the greatest impact on individuals or wider society."[15] The narrower focus on individuals rather than particular demographics, groups or society as a whole fails to reflect the nature of digital technologies which forge connections, groups, networks and communities. Ignoring this leaves vast numbers of users, including children and vulnerable people, exposed to manipulation, abuse and bullying at a worrying scale - both online and offline.

17. Facebook's internal analysis of its role in the US election noted that categorising electoral disinformation campaigns "as a *network* allowed [them] to understand coordination in the movement and how harm persisted at the network level. This harm was more than the sum of its parts." The report went on to conclude:

> *Because we were looking at each entity individually, rather than as a cohesive movement, we were only able to take down individual Groups and Pages once they exceeded a violation threshold. We were not able to act on simple objects like posts and comments because they individually tended not to violate, even if they were surrounded by hate, violence, and misinformation.*[16]

18. **The Online Safety Bill should account for the fundamental networking principles of online platforms, which promote connections and groups, and tackle harm at a much broader level.**

---

[12] Director General Ken McCallum gives annual threat update 2021
[13] Carbis Bay G7 Summit Communique (PDF, 430KB, 25 pages)
[14] The New Atlantic Charter 2021
[15] Online Harms White Paper - GOV.UK
[16] Facebook Stopped Employees From Reading An Internal Report About Its Role In The Insurrection. You Can Read It Here.

Are there any contested inclusions, tensions or contradictions in the draft Bill that need to be more carefully considered before the final Bill is put to Parliament?

Journalistic content and content of democratic importance

19. Not only does the Bill have weak provisions for tackling disinformation, it also **includes clauses which may actually legitimise harmful content**. Clauses 13 and 14 create specific duties for journalistic content and political debate. They are underpinned by definitions in Clause 39 and 40. Collectively, these **carve outs create loopholes whereby bad actors can create or share harmful content which will be protected on the grounds of newsworthiness**.

20. The definition of "news publisher content" includes news content and commentary as well as "gossip about celebrities, other public figures or other persons in the news". However, **the definition of "news publisher" is sufficiently broad as to potentially include anyone who sets up an eligible news website in the UK**. The bar for entry is extremely low, and **would allow bad faith actors** to circumvent online safety duties.[17]

21. Particularly worrying is that the exemption extends to when "a link to a full article or written item originally published by a recognised news publisher" is posted on a Category 1 service (13.10.iii). This may mean that **any posts on social media which include a link to a news site are exempt from services' safety duties,** opening up a whole host of **worrying scenarios permitting news content to be misrepresented and manipulated without recourse.**

22. **How "journalistic content" (Clause 14) differs from "news publisher content" is unclear.** Why include provisions for journalistic content if news content is exempt from the regime? This may be to account for journalists posting views and opinions on platforms directly rather than via the news sites (i.e. Tweeting live opinions or facts). Who qualifies as a journalist is also vague (e.g. is a Government Minister who practiced as a journalist before being elected included?) and again risks being **a loophole for bad actors to post harmful content under journalistic pretences**.

23. Another layer of worrying provisions, as regards disinformation and democratic harms, are the **carve-outs for political debate** or "content of democratic importance" (Clause 13). This states that Category 1 services must use "systems and processes" to **ensure that the "democratic importance" of content is considered in moderation decisions**. T&Cs must reflect these considerations. Content in this category includes news publisher content as well as content that "is or appears to be, specifically intended to contribute to democratic political debate in the UK or in any part or area of the UK". **This definition is vague and broad, raising many questions about what constitutes**

[17] Online Safety Bill: Five thoughts on its impact on journalism, LSE Media Blog, June 20201

legitimate political speech. The Explanatory Notes accompanying the Bill state "such content would be content promoting or opposing government policy and content promoting or opposing a political party", raising further questions about where the harm thresholds sit.[18]

24. It may be, for example, that hateful content targeted at political candidates is given special treatment on the grounds that it is "intended to contribute to democratic political debate". Such abuse is already particularly acute for female MPs. No female MP who was active on Twitter during the 2017 Election was free from online intimidation. During the election, Black and Asian women MPs – representing only 11% of all women in Westminster at that time – received 35% more abusive tweets than white women MPs.[19] This in turn has democratic consequences, affecting the ability of women MPs to fulfil their mandate safely and, at times, deterring them from (re-)running for office. **Any language in the Bill which inadvertently legitimises hateful content needs to be reworded.**

25. The result of these provisions is that fake news sites could be deemed out of scope of the regulations due to such a broad definition of "recognised news publisher" (40); radical views shared by extremists may be granted additional protections based on their supposed contribution to "democratic political debate" (13.6.b) ; anyone could misrepresent the contents of a news article when linking to it via social media (39.10.iii). These are **serious unintended consequences of the freedom of speech provisions, which are crucial to protecting fundamental rights but should not give disinformation a free pass. We believe that the language in Clauses 12, 13, 14, 39 and 40 must be revisited to avoid legitimising disinformation campaigns by bad actors.**

Freedom of speech

26. **Free speech is a fundamental right which must be protected**. The initial duty of care concept in the Online Harms White Paper navigated free speech by proposing a broad systems based duty of care to online platforms. The draft Bill takes a different approach, with different duties of care for separate categories of content. This takes the Bill in a different direction - **further away from a focus on systems and more to a focus on content.** This is particularly true of "content that is harmful to adults", the management of which the draft Bill presently defers to company T&Cs.

27. Where content is legal but harmful, **the Bill should focus on tackling the *reach* of content and not the speech itself. This should be managed via an amendment to Clause 11, as set out above and aligned with paragraph 241 in the House of Lords' Communication and Digital Committee's recent report, which demands platforms apply harm reduction by design measures[20].** This would reduce the virality of harmful

[18] Online Safety Bill: Explanatory Notes,
[19] Black and Asian women MPs abused more online, Amnesty International

content, introducing friction into an otherwise frictionless system, and avoid a takedown approach which deletes legal content. **Without such a change the Bill risks entrenching the status quo, where online platforms make and apply the rules.**

28. Some advocates of free speech are proposing tackling the free speech concerns raised by Clause 11 by removing "content that is harmful to adults" from the Bill and instead criminalising certain speech. This will shift harm reduction to a criminal investigation and prosecution regime. Harmful content will need to be reported to authorities, who will investigate and may ultimately take the case to court. This is a very protracted process when it takes place, putting an onus on victims to retread painful ground at great personal and financial cost. It is also widely acknowledged that such crimes are heavily underreported[21]. **Focusing on systems rather than content better protects freedom of expression while still reducing harm, and avoids further criminalising speech.**

29. **The powers reserved for the Secretary of State across the Bill are too broad and risk undermining both free speech and the independent regulatory regime.** There are unprecedented powers granted to the Secretary of State, such as Clause 33.1 which allows them to amend OFCOM's codes of practice to reflect government policy. These should be removed to ensure the independence of the regime.

Examples of harm reduction by design

Below are some examples of how technology companies have introduced design features to reduce the amplification of harmful content. These changes were made after intense public campaigns of shaming the platforms with open letters, petitions, media pressure and forcing Twitter and Facebook to meet victims of harm. This further underscores why we need regulation of legal but harmful content, for the government to require these design changes and be able to assess their impact. Otherwise they will continue to happen in a piecemeal way, with campaigners and researchers spending years and millions of pounds for tweaks that the platforms can make happen overnight if they were required to by law. As the below examples demonstrate, tech firms do indeed have the agility and insights to stem the spread of harmful material at pace and at scale.

---

[20] Free for all? Freedom of expression in the digital age

[21] Against Hate: Tackling hate crime in the UK, Amnesty 2017; How much online abuse is there?, The Alan Turing Institute, 2019

Read before you Tweet

Following a pilot in 2020, Twitter is planning to introduce a new design feature to encourage users to read articles before they retweet them, in an attempt to stem the flow of viral misinformation. In the pilot, Twitter prompted users who were about to retweet articles that they *hadn't* read to read the content before sharing. The result was that people opened the article 40% more often after seeing the prompt, and some people (Twitter hasn't disclosed the exact figure) didn't end up retweeting at all after reading. This is a huge shift in behaviour from a simple design intervention. The result is more informed users and a reduction in the virality of misinformation, harassment and hate speech.

Facebook's News Ecosystem Quality

In the days following the 2020 US Presidential election, misinformation about the election results flooded social media. In response, Facebook made a temporary change to its News Feed algorithm to give prominence to information from mainstream media outlets. To achieve this, Facebook dialled up the weighting of its "news ecosystem quality" (NEQ) score, a ranking Facebook assigns to news outlets based on signals about the quality of their journalism. According to internal sources at Facebook, the NEQ score usually plays a minor role in determining News Feed content, but concerns over the nature and scale of election disinformation drove senior executives including Mark Zuckerberg to temporarily increase NEQ's weighting. This resulted in a spike in visibility for mainstream news outlets.

This intervention is another example of how design choices can be made to reduce the reach of harmful material, as well as counter false information with that which is more verifiable. While it would undoubtedly be preferable for an independent regulator to determine which content is harmful, rather than a tech platform, this approach demonstrates that companies can respond at pace when focused on harm reduction, and that such design choices are already available to them.

"Break the Glass" measures by Facebook to slow the spread of electoral disinformation

In April 2021, BuzzFeed published an internal report by Facebook employees summarising the company's analysis of, and efforts to engage with, the social media fallout following the 2020 presidential election. The report explains how a taskforce was created to analyse and respond to electoral disinformation. In the report, members from the taskforce state:

> *We were also able to add friction to the evolution of harmful movements and coordination through Break the Glass measures (BTGs). We soft actioned Groups that users joined en masse after a group was disabled for PP or StS, this allowed us to inject*

friction at a critical moment to prevent growth of another alternative after PP was designated, when speed was critical. We were also able to add temporary feature limits to the actors engaging in coordinating behaviors, such as the super posters and super-inviters in the Groups that were removed, to prevent them from spreading the movement on other surfaces. These sets of temporary feature limits allowed us to put the breaks on growth during a critical moment, in order to slow the evolution of adversarial movements, and the development of new ones. Our ongoing work through the disaggregating networks taskforce will help us to make more nuanced calls about soft actions in the future in order to apply friction to harmful movements.

Removing direct messaging feature for under-16s

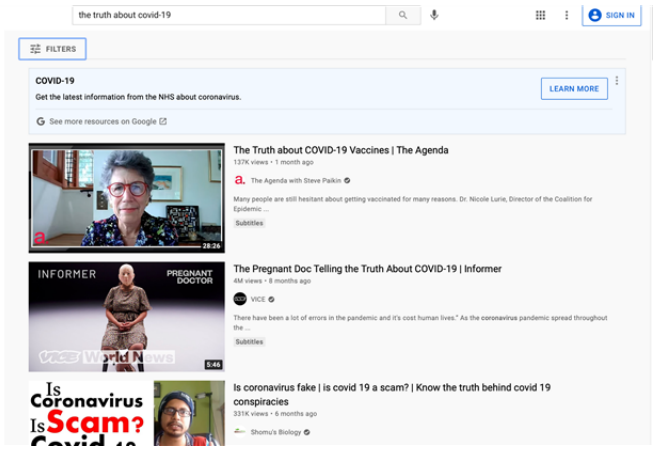
In April 2020, TikTok removed its direct messaging features for users under the age of 16 in an attempt to reduce the amount of grooming and bullying taking place in close conversations. It was the first time a major social-media platform has blocked private messaging by teenagers, on a global scale.

WhatsApp limits forwards

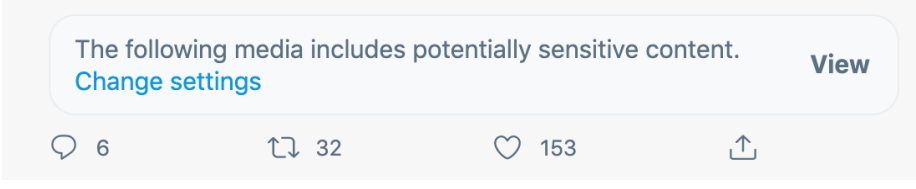
During the first wave of the Covid-19 pandemic, WhatsApp sought to address the “infodemic” by imposing a limit on message forwarding to slow the spread of mis and disinformation. Any frequently forwarded message: ie: forwarded more than five times, would get slowed down with users able to only forward that on to one user at a time.

YouTube search results on Covid

YouTube also took steps to address the massive spread of Covid-19 conspiracy videos on its site by prominently placing authoritative sources on the top of the search page. For instance, a search for “The truth about Covid-19” has a link at the top to an official NHS source.



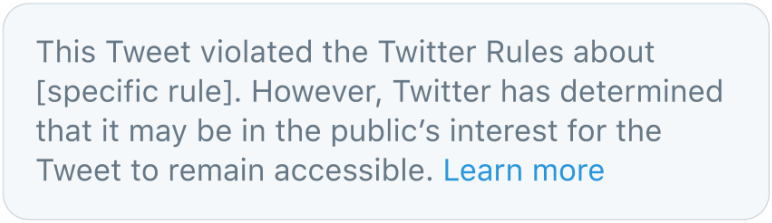
Warning messages for sensitive media



Most of the main platforms use warning messages to inform users about sensitive content. These messages are overlaid on specific Tweets or Posts, warning users about the nature of the content and requiring them to click through before they can view it. The messages stay on the site - content is not removed.

These are generally applied to content that has been marked (either by the person Tweeting it or following reports by other users) as “sensitive”, such as media that includes adult content, excessive gore or violence. This reduces the risk of users inadvertently witnessing content they might find harmful or distressing, but allows users who do want to find such content to access it. Users can choose whether to turn this feature on/off, so they don’t have to click through to view sensitive content.

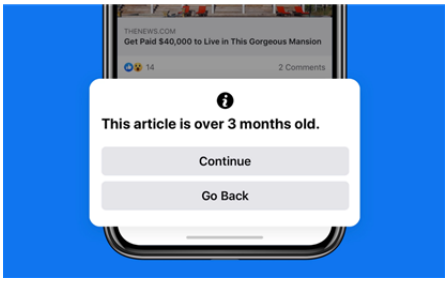
Twitter’s warning messages - public exemption policy media



In June 2020, Twitter applied for the first time its “public exemption policy”. The policy states that when a Tweet contains harmful content but is deemed to be in the public interest, the Tweet will be placed behind a notice. Such content would include harassment or hateful conduct, content which is in breach of Twitter’s T&Cs and for the majority of users would have to be taken down. Instead, in such instances, the notice would be applied which still “allows people to click through to see the Tweet” but “limits the ability to engage with the Tweet through likes, Retweets, or sharing on Twitter, and makes sure the Tweet isn't algorithmically recommended by Twitter”. This is an example of what it means to protect free speech while challenging unlimited reach. The exception only applies to elected or government officials with over 100,000 followers, and aims to “limit the Tweet’s reach while maintaining the public’s ability to view and discuss it”.

Facebook old story pop-up

In 2020, Facebook announced that it would introduce a notification screen warning users if they try to share content that’s more than 90 days old. They’ll be given the choice to “go back” or to click through if they’d still like to share the story knowing that it isn’t fresh.



Facebook acknowledged that old stories shared out of their original context play a role in spreading misinformation. The social media company said “news publishers in particular” have expressed concern about

old stories being recirculated as though they're breaking news.

Twitter - Harmful tweets prompt

In a recent blog post, Twitter announced the trial of a new product feature that temporarily autoblocks accounts using harmful language, such that they're stopped from being able to interact with a user's account. In the post, Twitter states:

> We are also continuing to roll out our replies prompts, which encourage people to revise their replies to Tweets when it looks like the language they use could be harmful. We found that, if prompted, 34% of people revised their initial reply or decided not to send their reply at all and, after being prompted once, people composed on average 11% fewer potentially harmful replies in the future.

About Reset

Reset (www.reset.tech) was launched in March 2020 by Luminate in partnership with the Sandler Foundation. Reset seeks to improve the way in which digital information markets are governed, regulated and ultimately how they serve the public. We will do this through new public policy across a variety of areas – including data privacy, competition, elections, content moderation, security, taxation and education.

To achieve our mission, we make contracts and grants to accelerate activity in countries where specific opportunities for change arise. We hope to develop and support a network of partners that will inform the public and advocate for policy change. We are already working with a wide variety of organizations in government, philanthropy, civil society, industry and academia.