Written evidence submitted by Reset (OSB0138)

Summary

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. It could not be a more timely piece of legislation.

2. The draft Bill is not perfect. The language is vague in many places. **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. Our proposals for the Bill are highlighted throughout this document and summarised below.

3. To succeed, **the Bill must tackle the design features and algorithms which amplify harm.** If it becomes a regime hinged 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. It must focus on tackling *reach* without curbing *speech*. This must be backed up with **powers to inspect the algorithms of services** to better understand the key drivers of harm.

4. We believe that the Bill must find a way to **tackle disinformation and its collective impact on society**. Despite regular observations by the Government and its international allies, including at a G7 level, of the threat disinformation poses to democracy, the drafting of the Bill makes it unclear as to whether disinformation would be in scope. In fact, **there are sections of the Bill which may even promote disinformation.**

5. The draft Bill **raises freedom of expression concerns, particularly in the powers granted to the Secretary of State.** Where the draft Bill does attempt to preserve freedom of expression, it must **ensure such protections apply to all users equally and do not give certain users special privileges to spread harm.**

6. The Bill **should include paid-for advertisements in scope to avoid the risk that bad actors buy their way out of the regulations**. Many of the harms which the Bill intends to cover are witnessed in ads as well as in user-generated-content. By excluding ads, the Bill creates incentives to shift harm into the paid advertising realm - keeping it on the platforms and perhaps even more prominent.

[1] Online Nation – 2020 report, p37, Ofcom, 2020

7. Appended to this submission are three documents which set out: a) How harm reduction by design can be applied in practice; b) an extended argument for broadening the definition of harm to account for collective impact; and c) case studies of disinformation. These lay out in more specific terms how the Bill can be improved and where it causes confusion.

8. Many countries are grappling with the same issues as the UK. There are similarities between the various regulatory approaches, as well as many differences. In our view, while the UK Bill certainly has the promise to lead the way globally, **there are elements of other international regulations - particularly the EU's Digital Services Act - which have an edge on the UK Bill.** Separate to this submission, we will submit an international comparison of online safety regulations in key jurisdictions.

Proposals and amendments

Content in scope

1. **Retain "adults' risk assessment duties"** (Clause 7)
2. **Amend Clause 11** to place a duty on companies to apply systems and processes to mitigate against the amplification and targeting of content that is harmful to adults.
3. **Remove 11.2** which defers management of content that is harmful to adults to the services' T&Cs.
4. **Encourage the regulator to include disinformation** as content that is harmful to adults.
5. **Tighten the definition of "recognised news publisher"** in Clause 40 to avoid creating a loophole for bad faith actors running intentionally misleading sites and outlets.
6. **Remove the exemption for links to articles in 39.10.iii** which creates a loophole allowing users to bypass regulations by embedding a link to a full article within user-generated content, even when that UGC misrepresents or has no relation to the linked article.
7. **Revisit the definition of "content of democratic importance"** to ensure it does not legitimise hate or disinformation.
8. **Include paid-for advertisements** as content in scope of the regulation.

Definition of harm

9. **Amend the definition of "content that is harmful to adults" in Clause 46 to include collective/societal harm**, reflecting language in the White Paper.

Enforcement

10. **Revise the language in Chapter 5 (Information Powers) to clarify that Ofcom has the power to audit the algorithms** of services in scope where appropriate. Ensure these powers can be applied based on concerns about all categories of content.
11. **Mandate that services share data with accredited academics** to improve research and policymaking.
12. **Give Ofcom the power** to respond to and push back on inadequate risk assessments.

13. **Give Ofcom the powers to enforce minimum standards** for compliance with the safety duties.
14. **Include civil society organizations i**n the disinformation advisory committee, and ensure the committee **looks at the harms caused by disinformation.**
15. **Remove the powers for the Secretary of State** to intervene in the regulator's agenda and enforcement.

Response

Algorithms and user agency

9. 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 reporting by the *Wall Street Journal* confirms that Facebook is well aware that its algorithm privileges sensational, polarising, and controversial content over all others -- and that they choose not to mitigate those harms in favour of high profits.[3] 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".[4] Similar investigations into Tiktok demonstrate the same outcomes; it's algorithm frequently drives users towards disturbing content.[5]

10. 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, otherwise known as "legal but harmful". Mark Zuckerberg elaborates on this in a blog post, stating that:

    "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."[6]

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

---

[2] Facebook Executives Shut Down Efforts to Make the Site Less Divisive, Wall Street Journal, 26 March 2020
[3] Facebook Tried to Make its Platform a Healthier Place, Wall Street Journal, 15 September 2021.
[4] YouTube Regrets, Mozilla Foundation, July 2021
[5] Inside TikTok's Algorithm, Wall Street Journal, 21 July 2021.
[6] A blueprint for governance and enforcement, blog post, Mark Zuckerberg, 2018 (updated 2021)

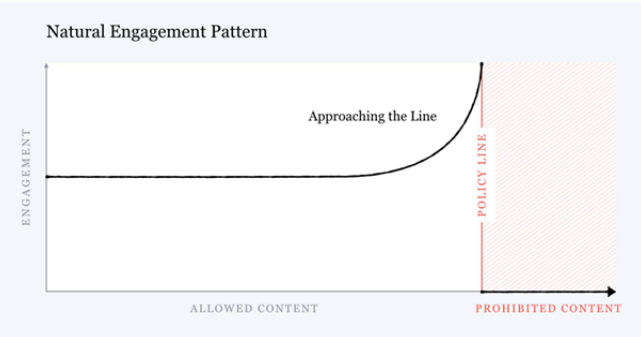


Fig. 1

Zuckerberg goes on to assert in this blog post (without evidence) that Facebook corrects for this human tendency to engage with extreme, disturbing, and provocative content by prioritising healthier conversations. Recent revelations from a whistleblower demonstrate that his public statements do not match his internal decisions. When presented with options for taming the artificial amplification of extremism in the NewsFeed, Zuckerberg himself rejected the changes so as not to reduce engagement, growth and revenue.[7]

11. **The Bill should be commended for including "legal but harmful" content as a category of harm**. It is the content which causes the exponential growth in Zuckerberg's engagement graph and which is being monetized at great societal cost. 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**.[8] Appallingly, this is true even when the harms include clear evidence documented inside the company that Instagram's focus on body imagery leads to a high incidence rate of mental health problems among teenage girls.[9] It bears repeating -- this content is not illegal. But the algorithmic curation of this content targeted at vulnerable audiences leads to unacceptable harms. **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'.**

12. **Encouragingly, the draft Bill acknowledges the power of design choices and algorithms in promoting harmful content.** The risk assessments for all categories of harm must account for the systems which promote harmful content, including algorithms; functionalities disseminating content; and how the design and operation of the service (including the business model) may influence risk. This is an important step in accepting the role of algorithms in promoting harm.

---

[7] Facebook Tried to Make its Platform a Healthier Place, Wall Street Journal, 15 September 2021.
[8] Facebook Struggles to Balance Civility and Growth, New York Times, 24 Nov 2020
[9] Facebook Knows Instagram is Toxic for Teenage Girls, Wall Street Journal, 14 September 2021.

13. The safety duties for content are more complex. For 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 however they choose as long as they set out their approach in the Terms and conditions ("T&Cs").

14. **Clause 11 means that even if services identify harms caused by their algorithms via their risk assessment duties, they are under no obligation to remedy those harms.** 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 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. It creates tiers of harm, with this category of content subject to the weakest harm reduction measures. And it gives platforms licence to continue to write their own rules.

15. The recent coverage in the *Wall Street Journal* of Facebook's secretive internal practices is further evidence of why platforms cannot be trusted to implement their own policies[10] . As a confidential, and previously unpublished, **report into Facebook's practices states: "We are not actually doing what we say we do publicly"**. **As currently drafted, the Bill would allow this to continue.**

16. **Rather than asking companies to write rules for *content*, Clause 11 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 strict limits on targeting**. **This should be supported by a Code of Practice on harm reduction by design drafted by Ofcom.** These measures would introduce friction into an otherwise frictionless system. They would **preserve *speech* while tackling *reach***. 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.

17. **It is important that support for "systems and process" does not become support for the use of *more* algorithms.** Algorithms have been proven to have an inherent

[10] Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt. Wall Street Journal, 13 September 2021.

discriminatory bias which penalises minorities. Simply algorithmically modifying content is not the solution. There must be more nuanced harm reduction methods in play and the Bill must avoid mandating the use of algorithms to tackle the harms they cause.

The role of Ofcom

18. **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.**

19. **Ofcom must be given clearer powers to inspect the algorithms and systems promoting all forms of harmful content.** The benefits of, and approaches to, algorithmic inspection are set out in this paper by Reset and Ada Lovelace Institute.[11] 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 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.** This authority should include not just access to data and code, but the power to query the systems engineers and product managers to determine the relative impact of design and features choices.

20. 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 and what impact it has on fundamental rights and the greater public good - such as public health, public safety or democratic culture. This in turn would help policymakers and the safety tech industry develop innovative ideas and products based on evidence and data. And it would **align the Bill with the Digital Services Act**, which does include this provision. **Just as the scientific community studies data associated with health care, climate change, and education, the impact of digital media on the public should be similarly evaluated by the collective work of experts.**

21. Current transparency arrangements with researchers remain at the whim of platforms, as the recent debacle with transparency researchers at New York University (NYU) demonstrates.[12] After consistently publishing world-leading, insightful analysis of Facebook's advertising practices, the company shut down the researchers' accounts to halt their progress. **Clearly, the transparency measures required by law should not be**

[11] Algorithms in social media: realistic routes to regulatory inspection
[12] We Research Misinformation on Facebook. It Just Disabled Our Accounts.

at the behest of platforms, which leave them susceptible to self-serving restrictions. As the NYU researchers wrote: "We believe that Facebook is using privacy as a pretext to squelch research that it considers inconvenient". The researchers had just begun work on important studies "to determine whether the platform is contributing to vaccine hesitancy and sowing distrust in elections. We were also trying to figure out what role the platform may have played leading up to the Capitol assault on Jan. 6." **This Bill needs to ensure transparency powers for researchers and academics are codified in law.** The platforms have demonstrated that they cannot be trusted to facilitate such transparency under the status quo.

22. **Ofcom must also be given the powers to push back on risk assessments which it deems to be inadequate**. As currently drafted, there is no penalty for companies writing poor risk assessments which fail, inadvertently or otherwise, to account for risk on their services. **Without this power, the regime is toothless.** Such a power is fundamental to this agenda and must be made available to the regulator as a priority. Large online platforms have a history of providing cosmetically appealing and substantively empty reports to government regulators.[13] **Authority to verify validity and penalise duplicity/inadequacy must be granted to the regulator to ensure compliance.**

23. In addition, **there must be minimum standards for compliance with the safety duties, perhaps through binding codes of practice**. At present, the codes produced by Ofcom will be non-binding and services will be able to argue that they are compliant with safety duties through other means if they choose not to adhere to the codes. In those instances, Ofcom can do very little to enforce against any breaches. **This again makes the regime toothless and affords a large amount of trust to platforms to "do the right thing" when time and time again they have done otherwise.**

Content in scope

Content that is harmful to adults

24. **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. 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 than the current**

[13] See, for example, this investigation by the Süddeutsche Zeitung examining Facebook's haphazard compliance with the German hate speech law.

draft requires. As per our recommendation above, we believe that "content that is harmful to adults" (Clause 11) should be managed by harm reduction by design measures rather than by company T&Cs.

Disinformation and collective harm

25. **Reset believes the Bill must meaningfully tackle disinformation.** The Government's commentary on the Bills 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.**

26. The UK has witnessed or been subject to multiple coordinated disinformation campaigns in recent years. **Online disinfo campaigns which spill into offline harassment of journalists[14]; state backed disinformation campaigns inauthentically amplifying partisan views on Scottish referendum[15]; climate change denial[16]; disinformation discrediting the Security Services' investigations into the Sergei Skripal poisoning[17]; and 5G conspiracy theories[18] are just some of the attempts to undermine trust in authorities and sow confusion**.

27. The Bill includes provisions for the creation of an advisory committee on disinformation (98). This committee **should include strong representation from civil society groups** in addition to representatives of UK users of regulated services and academic experts. It must be a wide forum to **ensure citizens affected by harmful misinformation are active participants** in the system and how it operates. The **committee should also have within its function an explicit remit for understanding the *harms caused* by disinformation,** and not just how disinformation is managed by services. A focus on harms, supported by input from victims and civil society, will strengthen the committee.

28. While the creation of this committee is welcome, it alone is insufficient in tackling a live and pervasive issue. **The EU's Digital Services Act tackles disinformation** by recognising that the use of 'VLOPs' (Very Large Online Platforms) poses 'systemic risks' to individuals and to societies. Article 26 of the proposed Act requires platforms to carry out risk assessments on systemic risks which include not just the dissemination of illegal content but also of:

---

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

intentional manipulation of their service, including by means of ***inauthentic use or*** ***automa***

29. In defining this category of risk, the EU notes that this concerns '*the intentional and, oftentimes,* ***coordinated manipulation of the platform's service****, with a foreseeable impact on health, civic discourse, electoral processes, public security and protection of minors, having regard to the need to safeguard public order, protect privacy and fight fraudulent and deceptive commercial practices."*

30. Experts and civil society groups are working with the EU to enhance this further, ensuring that it specifies the risks relating to disinformation and not just the 'intentional manipulation' of the platforms, and that the Act imposes further obligations on platforms to tackle these risks.

31. **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.[19] At an international level, the recent communique following **the G7 Summit committed G7 nations to "strengthening the G7 Rapid Response Mechanism to counter foreign threats to democracy including disinformation"[20]**; 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"[21]. 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.**

32. 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."[22] 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.**

---

[19] Director General Ken McCallum gives annual threat update 2021
[20] Carbis Bay G7 Summit Communique (PDF, 430KB, 25 pages)
[21] The New Atlantic Charter 2021
[22] Online Harms White Paper - GOV.UK

33. Facebook's internal analysis of its role in the US 2020 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.*[23]

34. Under current proposals, **the Online Safety Bill would do little to avoid a scenario like the violent fallout of the US Presidential election happening in the UK.** If the Bill is to keep people safe online as well as offline, and protect society at large, **it must account for the fundamental networking principles of online platforms, which promote connections and groups, and tackle harm at a much broader level.** This framework will permit oversight to account for the distortion of information markets caused by algorithmic curation that tends to increase the frequency (and normalisation) of extreme views, conspiracy theories, and other forms of harmful content.

Journalistic content and content of democratic importance

35. Not only does the Bill have weak provisions for tackling disinformation, it also **includes clauses which may actually legitimise disinformation and other 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**.

36. 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, including blog posts and sites.** The bar for entry is extremely low, and **would allow bad faith actors** to circumvent online safety duties.[24]

37. 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

---

[23] Facebook Stopped Employees From Reading An Internal Report About Its Role In The Insurrection. You Can Read It Here.

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

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.**

38. **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 former journalist included once they have left the vocation?) and again risks being **a loophole for bad actors to post harmful content under journalistic pretences**. In addition, **such vague definitions will permit the companies to dodge obligations for risk mitigation by citing the red lines around this kind of content; and it may well chill regulators from taking legitimate actions that might be perceived in tension with these broad exemptions.**

39. 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 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.[25]

40. 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.[26] 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] Online Safety Bill: Explanatory Notes,
[26] Black and Asian women MPs abused more online, Amnesty International

41. 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.**

Paid-for advertisements

42. **Harmful content can appear in advertisements as well as in organic UGC**. However the draft Bill excludes paid ads from scope. This means that **those intent on spreading harmful content can simply buy their way out of the regulations. The same content posted on social media by a user, and therefore subject to the safety duties of the Bill, would be able to circulate freely if placed in an advertisement.**

43. There are **many examples of harmful content being targeted at users via ads.** This year, Reset Australia ran an experiment to see whether ads promoting smoking, gambling, alcohol and extreme weight loss could be targeted at under 18s.[27] All the ads were approved by Facebook, although ultimately not published by Reset. Similar research by Global Witness secured Facebook's approval for ads inciting sectarian violence in Northern Ireland. [28] The ads were also not published. However, in the 2016 US Presidential Election, Donald Trump's campaign team targeted real ads at voters, in some cases with the aim of suppressing their vote.[29]

44. **To avoid commercialising harmful content even further, the Bill must include in scope paid-for advertisements.** Without this amendment, the Bill risks rerouting more money into the big tech companies and the digital advertising industry.

Appendix 1 - 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

[27] Facebook: Smoking and alcohol ads 'target Australian children'.
[28] The Big Tech business model poses a threat to democracy
[29] Revealed: Trump campaign strategy to deter millions of Black Americans from voting in 2016

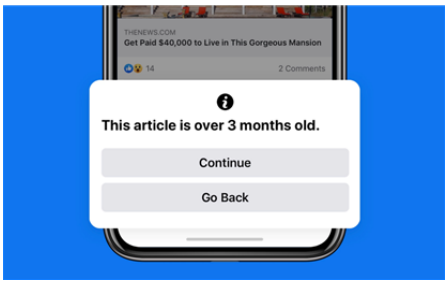
campaigns of shaming the platforms with open letters, petitions, media pressure and forcing Twitter and Facebook to meet victims of harm. This further underscores why we need regulation of legal but harmful content, for the government to require these design changes and be able to assess their impact. 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.

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 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.

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

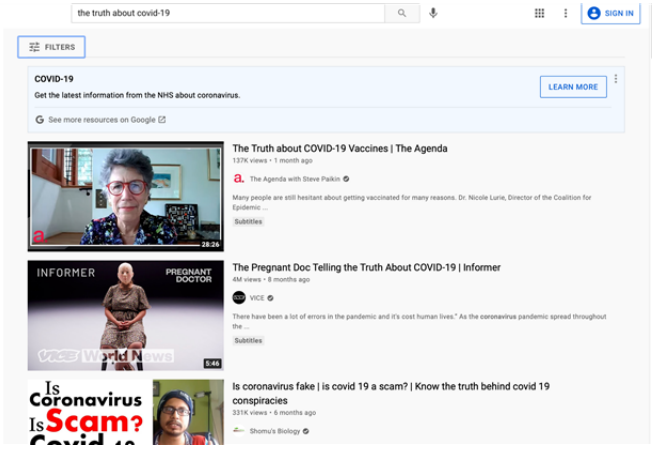
In April 2020, TikTok removed its direct messaging features for users under the age of 16 in an attempt to reduce the amount of grooming and bullying taking place in close conversations. It was the first time a major social-media platform has blocked private messaging by teenagers, on a global scale.

WhatsApp limits forwards

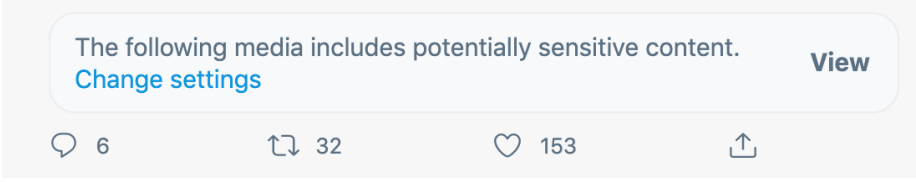
During the first wave of the Covid-19 pandemic, WhatsApp sought to address the "infodemic" by imposing a limit on message forwarding to slow the spread of mis and disinformation. Any frequently forwarded message: ie: forwarded more than five times, would get slowed down with users able to only forward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

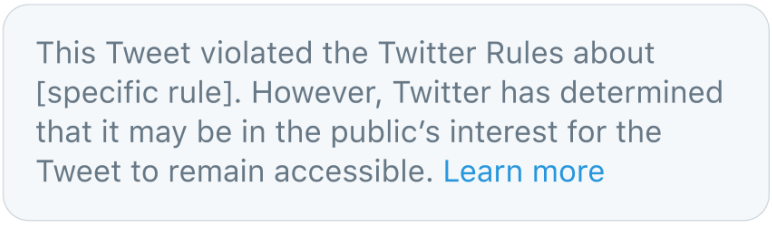


Most of the main platforms use warning messages to inform users about sensitive content. These messages are overlaid on specific Tweets or Posts, warning users about the

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".

Reset's accuracy prompts test

Reset has worked with two leading experts at MIT and University of Regina in Canada to pursue an empirically grounded strategy to fight mis and disinformation. Their academic research distilled in this paper, Lazy not biased, found that most **people actually share misinformation out of laziness and lack of friction in platform design, rather than inherent bias.** They tested whether a simple reminder about the importance of accuracy in the social media feed (content neutral and non-partisan nudges or "accuracy prompts") could yield an increase in cognitive discernment of true/false content to deter a significant percentage of harmful sharing of disinformation.

Using a random sample of 3000 likely US voters (studies were also carried out in France, Italy and Canada with similar results), they found that the "accuracy prompt" Facebook ad shows the potential to change the proportion of true and false content shared on social media by 7-9%. (This testing was done via YouGov). That may seem a modest change but for highly vulnerable groups (eg. those that share disinformation with great frequency) we are moving from a baseline of users that share disinformation as often or even more often as they do accurate information to a 7% advantage for accurate information. This represents as much as a 14% decrease in disinformation sharing.

Appendix 2 - The case for tackling collective harm

The Online Safety Bill is a pioneering piece of legislation which aims to improve online safety for individual users, in particular for children. The companies in scope are predominantly the Big Tech platforms where the vast majority of harmful content is posted and witnessed. These platforms offer users unprecedented connectivity to friends, family, public figures, strangers, authorities and much more. They exist to bind us together and forge connections.

These are some of the wealthiest firms in the history of the world, whose primary source of revenue is advertising. The more time people spend online, the more ads they can sell. It is in their interest to keep users logged-on so that they stay clicking. To do this, platforms serve up engaging content and that which most grabs our attention tends to be sensationalist, subversive and controversial. Algorithms, which are trained to drive engagement at the expense of nothing else, push content and network recommendations which we wouldn't necessarily seek out or find ourselves, but which we can't resist. For example, an internal report by **Facebook in 2016 found that 64% of all extremist group joins on the platform were due to recommendation algorithms.**[30]

The attraction to sensationalist content is nothing new. It's human nature. But in our digital world, and on large technology platforms, the pace and scale at which such content is witnessed is unprecedented. We are encouraged to push this content - be it on Pages, comments, stories or posts - with and beyond our network, while algorithms work furiously to secure engagement. We are rarely encouraged to assess the content we share for accuracy or potential harm, if we are encouraged to read it *at all*. **This creates false levels of trust in content, which is given more credence by virtue of being shared by trusted contacts.** Groups allow people to share perspectives with huge audiences with a single click, allowing new levels of reach which are impossible for individuals to replicate offline. Third parties can interact with, or insert themselves into, these networks with ease. The frictionless system ensures content spreads at speed and scale, feeding the recommendation algorithms with as much data as possible in order to improve targeted content**. This perpetuates a vicious cycle in which the individual becomes indistinguishable from their network.**

---

[30] Facebook Executives Shut Down Efforts to Make the Site Less Divisive

Collective harm

The result is an environment where **harmful content becomes normalised - where what was once at the fringes becomes mainstream, dominating Feeds and Posts.** Covid disinformation brought that into sharp relief. Research by the Institute for Strategic Dialogue (ISD) showed that Covid disinformation had twelve times the volume of interactions on disinformation sites as credible sites like the WHO and CDC.[31] **When factual information relating to a highly contagious virus is drowned out by disinformation, this becomes a *collective* harm.** It damages society at large, through the erosion of trust in official bodies and the growth of a culture where conspiracy theories are mainstream, as well as individuals. Separate research by ISD showed that false claims of voter fraud in the US Presidential election were shared at twice the rate of the best performing official pieces of information.[32] The resulting confusion and aggression this stirred up played out in a very public way. 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."[33]

Similarly, with the recent sexual abuse scandal in UK schools, the Everyone's Invited movement exposed how attitudes towards girls are being shaped by the normalistion of pornography and the volume at which pornographic material is viewed online.[34] Authorities have had to remind teenage boys that the pornographic material they watch online is fictional. **When such boundaries become blurred, the result is both a very real harm for individual victims as well as a broader collective harm to society.**

These are not isolated examples. Racist abuse, misogyny, climate change denial, homophobia, bullying, pro-suicide and self-harm and many other types of harmful content are as readily promoted to the widest possible audience, propagating harm at a network level. Often, this promotion of harmful material is coordinated by malicious groups with malign intent, as it is now far easier to target, access and influence groups online than it is offline. But just as problematic is how the business model of Big Tech prioritises content which will *spread* over that which is safe, nudging users to engage with harmful content at any cost. **Harms to society often stem from harms to individuals, and vice versa. One cannot be considered distinct from the other.**

Online Safety Bill

The definition of harm in the draft Bill is limited to content which has "a significant adverse physical or psychological impact" on an individual of "ordinary sensibilities".[35] This differs from the language in the Online Harms White Paper which proposed "prioritising regulatory action to

---

[31] Public Health Disinformation - ISD
[32] Narratives of Violence around the 2020 Presidential Election - ISD
[33] Facebook Stopped Employees From Reading An Internal Report About Its Role In The Insurrection. You Can Read It Here., Buzzfeed, April 2021
[34] https://www.bbc.co.uk/news/uk-56558487
[35] Draft Online Safety Bill

tackle harms that have the greatest impact on individuals or wider society."[36] 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. It means platforms will continue to distort public discourse, as well as long-held social and cultural norms. **The Online Safety Bill should account for the fundamental networking principles of online platforms, which promote connections and groups, and ensure that any legislation prevents harm at a much broader level.**

Appendix 3 - Disinformation case studies

Below are a number of real world examples of disinformation which would not cross the illegality threshold and so would only be in scope of the OSB if they met the definition of "content that is harmful to adults". This means, to be in scope, "the nature of the content is such that there is a material risk of the content having, or indirectly having, a significant adverse physical or psychological impact on an individual of ordinary sensibilities".

It is unclear which of the below would reach that threshold. It could certainly be argued that all of these result in direct or indirect harm, either physical or psychological, to individuals. Many of them have a direct societal harm, however of course that is not included in the definition at present. It is a useful exercise to demonstrate where the definition creates uncertainty and where serious instances of disinformation may be free to circulate online without recourse.

[36] Online Harms White Paper - GOV.UK

	Headline	Example	Comments
1	State backed disinformation campaigns inauthentically amplifying partisan views on Scottish referendum	Twitter and Facebook have identified swathes of fake accounts linked to the governments of Russia and Iran which amplify messages from pro-independence campaigners. Facebook has also identified, and deleted, a page called Free Scotland 2014 which was traced back to Iran and linked to fake news sites.	Unclear whether such campaigns would be captured under the current definition. Could be argued that encountering such information directly impacts views and outlooks; as well as indirectly pro-Unionists by damaging their cause and, by extension, the Union. Unlikely to qualify as "significant".
2	Coordinated disinformation campaign against journalists.	This example of a disnfo campaign against an AP journalist, which resulted in her losing her job, is just one of the rising number of cases of journos being subject to disinfo campaigns. In this instance, the journalist was targeted by a political group who mischaracterized her old tweets to portray her as antisemitic. She lost her job as a result. Recently, Christina Lamb, Chief Foreign Correspondent for The Times, found herself in a similar situation but hept her job. And the reporter Jon Snow	Disinfo campaigns against journalists are on the rise. Responses are generally a matter of internal policy for publishers but it is increasingly difficult for news outlets to distinguish between legitimate public reaction and coordinated attacks. It is unclear whether journalists are protected against such targeted campaigns by the OSB, and how the protections for journalists marry with the definition of harm. The attacks on Jon Snow and Nicholas Watt are examples of how online campaigns translate into offline harm.

		was physically assaulted by antivax campaigners who had encountered conspiracy theories online.	
3	Climate change	There are plenty of examples of disinformation campaigns sowing confusion and outright lies online about scientifically proven climate change realities. This includes false assertions about the evidence proving climate change, conspiracy theories about politicians calling for action and fake statistics about the environmental impact.	False information about climate change directly and indirectly impacts behaviour, health and wellbeing. In the medium to long term, the impact is highly significant and adverse.
4	The role of coordinated disinformation campaigns in loss of life and livelihood	The British co-founder of the White Helmets, a Syrian civil defence group, committed suicide following disinformation campaigns run by Russian and Syrian propagandists claiming the White Helmets faked evidence of Syrian atrocities.	An example of how a deluge of disinformation can have a direct and indirect detrimental psychological impact.
5	Islamophobic conspiracy theories (Eurabia and "No Go Zones") inciting violence and racial tensions. Anders Brevik, who carried out atrocities in Norway, was a follower of such beliefs.	Widely debunked far-right conspiracy theories about Islam run rife on social media and news media sites/blogs. Examples include a deal between Western governments and the Middle East to exchange Western sovereignty for oil; a plot by Islamic nations to take over Europe to create "Eurabia";	Another national security issue linked to disinformation and fake news. Issues around how far-right conspiracy sites are categorised vis a vis "news publishers" and whether their content would be deemed out of scope.

		claims of “No Go Zones” in Western nations which are run by Sharia law and bar non-Muslims and police.	
6	Families, children and conspiracy theories	Sites and movements such as QAnon risk having an indirect adverse effect on families whose members start to believe anti-vax disinformation or wild conspiracy theories about e.g. the political elite. There are increasing examples of families torn apart by such movements, with the impact on children whose parents start to distrust authorities and remove their children from educational services.	While certainly indirect harm, the threshold of significance makes it unclear whether this would be in scope. Such disinformation has a long-tail, with the impact of children being removed from educational and healthcare services having a lasting - but perhaps not immediate - effect.
7	Manipulated video image of political candidates	Deep Fake Video of Speaker Nancy Pelosi, appears to show her slurring and impaired. The video was viewed more than 2.5 million times on Facebook and Facebook refused to take it down, despite clearly being a deep fake. A deepfake video of Boris Johnson endorsing Jeremy Corbyn, created for academic and campaigning purposes, shows how else this technology can be applied.	An example of how deep fakes can be used to manipulate videos and messages. More advanced technology could have more severe implications.
8	Disinformation about the	In addition to the racial abuse	

	sexuality of footballers	received by footballers and their families, many are also on the receiving end of disinformation about their sexuality. Most recently, Premier League player James Rodriguez was the subject of false rumours claiming that he had not played because he had undergone gender reassignment surgery. The false information involved people mocking up images saying he would never play again.	
9	Coordinated campaigns to stir-up racial tensions ahead of the 2016 presidential election	According to the US Senate Intelligence Committee, disinformation campaigns by Russia in the US 2016 Presidential election targeted African-American communities more than any other group. "By far, race and related issues were the preferred target of the information warfare campaign designed to divide the country in 2016". This included false claims about police brutality, political party funding (such as Hillary Clinton received $ from the Klu Klux Klan) and about the BLM movement.	Disinfo campaigns by state actors are far more nuanced than simply targeting individuals or groups of individuals with abusive content. In this case, the aim was to sow confusion and division in order to subtly change opinions and reduce voter engagement. This has major democratic implications. It could be argued that such a *campaign* fits the definition of harm, but it is the nature and scale of the content rather than the content itself which causes the harm. May be out of scope.

10	Coordinated response to discredit Sergei Skripal poisoning	Campaign by Russian officials and others calling into question the authenticity of images published by HMG, and promoting #skripalhoax	Disinformation had no bearing on the poisoning itself but rather on post hoc public perception of the event and whether/how it took place. Weakens trust in security services and sows confusion, but does not clearly fulfil the definition.
11	US electoral fraud	Perhaps the most high profile conspiracy in modern democracies is the evidence-free claim that the 2020 elections were stolen and the results therefore illegitimate. This conspiracy is widely disseminated on social media and echoed by mainstream media and prominent political leaders in whole or in part. As a result, a substantial percentage of the electorate believes the election was fraudulent and has lost confidence in electoral institutions.	This is a particularly pernicious example because the disinformation is centrally about politics and elections in the most consequential way possible. Would it be subject to protection due to "democratic importance"? The platforms have all changed terms of service to prohibit election delegitimization; but they are failing to achieve that objective. Should the regulator insist they do? Or insist they shouldn't? Consider the possibility that this scenario occurs in a future Scottish referendum.
13	UN Global Compact on Migration	In December 2018, The UN Global Compact on Migration was subjected to a wave of misinformation from over 50,000 Twitter users in multiple languages, as well as manipulation of	Given this has evidently resulted in indirect physical impact (the Christchurch massacre) it feels this should qualify. However it's unlikely such campaigns would be picked up before or unless a mass murder took place.

		the YouTube recommendation algorithm by far-right outlets. The misleading content pushed the far right Great Replacement Theory and spread lies about the outcomes of the Compact. Countries including Brazil, the USA, Austria, Bulgaria, the Czech Republic, Slovakia, Poland, Australia and Israel pulled out. Later, the barrel of the Christchurch killer's gun was found with the chilling inscription: 'Here's your global compact on migration.'	

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.

27 September 2021