Exemptions, exceptions and exclusions: Why the Online Safety Bill protects disinformation and abuse over freedom of speech and journalism

Prepared by:
Kyle Taylor, Fair Vote UK (kyle@fairvote.uk)
and
Ellen Judson, Demos (ellen.judson@demos.co.uk)
with contributions from:
The Global Project Against Hate & Extremism

EXECUTIVE SUMMARY

The UK's Online Safety Bill has been lauded by the government as "world-leading," with claims that it will make the UK the "safest place in the world" to go online. Major loopholes in the bill, however, mean **this bill represents online safety in name only.**

Moreover, the UK has just become a signatory to the Declaration for the Future of the Internet. Among other things, the declaration affirms the UK's commitment to 'ensure that the Internet reinforces democratic principles and human rights and fundamental freedoms' and that 'the well-being of all individuals [is] protected and promoted', and to 'promote safe and equitable use of the Internet for everyone, without discrimination', encouraging 'pluralism without fear of censorship, harassment or intimidation.'

As currently drafted, the Online Safety Bill **fundamentally fails to meet the ambitions set out in these principles.**

Protecting Freedom of Expression

It is right that there are significant protections for freedom of expression in a Bill which will affect how people communicate on the internet. Given that the Bill places duties on platforms to take certain actions regarding content, including taking it down, demoting it or reducing access to it, **there is a clear need for freedom of expression safeguards**, to reduce the incentives for platforms to vastly over moderate in their attempts at compliance. However, it is crucial that these safeguards do not themselves either incentivise under moderation, or provide a loophole through which platforms can simply continue driving engagement for profit - which would fundamentally undermine the purpose of the Bill - to make the online world safer.

Exemptions, exceptions and exclusions that undermine freedom of expression

Despite these safeguards, however, the government has also written into the bill additional exemptions, exceptions and exclusions, on the basis of seeking to better protect freedom of expression. Far from protecting freedom of expression, however, this means **freedom of expression safeguards are not a level playing field**, and means that already privileged entities, such as the press and politicians, are likely to be afforded greater freedoms by platforms than ordinary people. **These measures are:**

A media exemption which, as drafted, means platforms are exempt from having to take any action around safety duties relating to news publishers' content *even if that content is illegal or demonstrably harmful for children*. The criteria to qualify as a news publisher which are laid out in the bill are **arbitrary and very easily met**, meaning a bad actor - foreign or domestic - could easily qualify as media and manipulate the rules. **Exempting media content from a systems-based approach is not coherent.** If a platform has a truly effective and proportionate system to reduce the risk of harms to users, as required in the Bill, media content will only be negatively and disproportionately affected by that system if and where it poses a significant risk of harm to users.

Protections for 'journalistic' content that require platforms to take particular care over how their content moderation processes (including action taken against content or users sharing content) might interfere with the free expression of content that is published for the purposes of journalism (including both news publisher content and regulated content). This includes an expedited complaints procedure. **This moves towards creating a two-tier system without appropriate protections**: It is suggesting that content which is deemed journalistic, regardless of what it is, whether or not it is accurate, could cause harm, or is relevant to a debate - will benefit from additional protections.

Protections for 'democratically important' content that require platforms to take particular care over how their content moderation processes (including action taken against content or users sharing content) might interfere with the free expression of content that is or appears to be specifically intended to contribute to democratic political debate in the United Kingdom (including both news publisher content and regulated content). This is a category which is either so wide as to be meaningless, or privileges speech of those discussing government policy and political parties - likely to disproportionately be politicians themselves. The government has made clear that preserving freedom of expression is the most important duty, meaning that other protections outlined in the bill, like safety from abuse, are subjugated if the topic is of "democratic importance." This *also* applies with regard to how platforms treat content which is harmful to children.

The bill will lead to platforms effectively making decisions based on who a user is rather than the potential for harm, particularly at scale, blatantly privileging certain users over others and creating a two-tiered Internet. This is unequal and dangerous.

Harm is harm wherever it appears online and mitigation measures to remove or slow the spread of such content should be universally applied based on its likelihood to cause harm. Content online does not somehow become less harmful depending on what user posted it or whether a topic is deemed currently relevant to politics.

Paid ads are almost entirely out of scope, meaning users and entities can pay to evade the rules. **The content is no less harmful because it has been paid for.** In many cases, the opposite is true, *and* the platforms are directly profiting. Adverts spreading vaccine disinformation, for instance, would be excluded from what platforms would be expected to mitigate the harm of.

Around the world, it is often political leaders, news publishers and paid advertisers – those who are most likely to be protected under the Bill – that can cause the most harm on social media due to their disproportionate influence and reach. As drafted, the bill exempts, excludes or allows exceptions for what extensive research around the world has shown are the most common sources of harmful content online. **This harms, rather than helps, a free and independent press.** Taking action against abuse and disinformation is not antithetical to journalism - it *supports* journalism. **Maria Ressa, Nobel Peace Prize winner and journalist noted that** "It can only protect media freedom if it is able to cut off disinformation upstream. Creating [a] special media redressal mechanism may sound good but will ENABLE industrial scale disinformation'.

Inspiring Terrorism: The Real-World Impact of These Loopholes

A recent example - the white supremacist terrorist murder of ten people of colour in Buffalo, New York, USA brings to life the convergence of all the bill's loopholes:

- **The media exemption** would have allowed Tucker Carlson of Fox News, a recognised news publisher in the United States to endorse the dangerous white supremacist conspiracy theory called the "great replacement", in which ethnic minorities are supposedly replacing white Europeans in order to further a left-wing political agenda that inspired the mass-shooting on May 14th, with the shooter acknowledging he had absorbed such content from Carlson.
- **The journalistic content exception** would have allowed these theories to be promoted by smaller-scale "citizen journalist" commentators online.
- Additionally, the "great replacement theory" has been widely promoted online by representatives of the US government, including the number 3 ranking Republican in the United States Congress. This would be protected by **the "speech of democratic importance" exception.**
- Lastly, politicians in the United States used paid ads to promote the great replacement theory. **These political ads would be out of scope for this bill**, yet contributed directly to real-world violence.

Far from reducing the risk of over moderation, these exemptions, exceptions and exclusions create perverse incentives for platforms: to over moderate the free expression of normal people who are more likely to fall outside these exemptions and are apparently, less deserving of freedom of expression protections, while allowing potentially hateful and abusive content to proliferate as long as it comes from people and entities that fall into these categories.

Recommended Amendment Areas

The bill should be amended to strengthen the protections for freedom of expression. This would allow for **further amendments** to remove the media exemption and special consideration for democratic and journalistic content as well as bring all paid ads explicitly in scope. We set out how this can be achieved in the attached briefing.

FULL BRIEFING

Overview

The Online Safety Bill, which had its 2nd reading in the UK Parliament on 19 April 2022, has been lauded by the government as "world-leading," with claims that it will make the UK the "safest place in the world" to go online. Major loopholes in the bill, however, mean **this bill represents online safety in name only.**

Moreover, the UK has just become a signatory to the Declaration for the Future of the Internet. Among other things, the declaration affirms the UK's commitment to 'ensure that the Internet reinforces democratic principles and human rights and fundamental freedoms' and that 'the well-being of all individuals [is] protected and promoted', and to 'promote safe and equitable use of the Internet for everyone, without discrimination', encouraging 'pluralism without fear of censorship, harassment or intimidation.'

As currently drafted, the Online Safety Bill **fundamentally fails to meet the ambitions set out in these principles.**

How does the Bill try to protect freedom of expression?

It is right that there are significant protections for freedom of expression in a Bill which will affect how people communicate on the internet. Given that the Bill places duties on platforms to take certain actions regarding content, including taking it down, demoting it or reducing access to it, there is a clear need for freedom of expression safeguards, to reduce the incentives for platforms to vastly over moderate in their attempts at compliance.

However, it is crucial that these safeguards do not themselves either incentivise under moderation, or provide a loophole through which platforms can simply continue driving engagement for profit - which would fundamentally undermine the purpose of the Bill.

The Bill rightly sets out clear freedom of expression safeguards. These are:

- A duty to have regard to the importance of protecting users' right to freedom of expression within the law.
- A duty to have a complaints procedure which enables users to complain if platforms are not complying with this duty, if their content or their use of the site has been affected
- A duty for platforms who take action against content that is legal but harmful to adults to clearly specify their policies in their terms of service and enforce them consistently
- A supercomplaint mechanism through which complaints can be made to the regulator if platforms are significantly adversely affecting the right to freedom of expression within the law of users of the services or members of the public

Despite these safeguards, however, the government has also written into the bill additional exemptions, exceptions and exclusions, on the basis of seeking to better protect freedom of expression.

Far from protecting freedom of expression, however, this means **freedom of expression safeguards are not a level playing field**, and means that already privileged entities, such as the press and politicians, are likely to be afforded greater freedoms by platforms than ordinary people. They are also likely to result in some cases in *preventing* platforms from taking even the minimal actions to keep users safe which platforms currently have in place.

Exemptions, Exceptions and Exclusions

A media exemption

As drafted, the media exemption means platforms are exempt from having to take any action around safety duties relating to news publishers' content *even if that content is illegal or demonstrably harmful for children*. News publishers also do not have to take any action on harmful user-generated content on their own sites (e.g. comment sections).

Protections for 'journalistic' content

Platforms must take particular care over how their content moderation processes (including action taken against content or users sharing content) might interfere with the free expression of content that is published for the purposes of journalism (including both news publisher content and regulated content).

They must specify how they will identify journalistic content and how they will protect the free expression of journalistic content above and beyond how they treat content generally.

They must also have an expedited complaints procedure available for 'a person who considers the content to be journalistic content' and has created, uploaded, generated or shared the content.

Protections for 'democratically important' content

Platforms must take particular care over how their content moderation processes (including action taken against content or users sharing content) might interfere with the free expression of content that is or appears to be specifically intended to contribute to democratic political debate in the United Kingdom (including both news publisher content and regulated content).

They must specify how they will identify journalistic content and how they will protect the free expression of journalistic content above and beyond how they treat content generally.

Platforms must specify how they will treat democratically important content above and beyond how they treat content generally, and ensure they apply this duty to a diversity of political opinion.

In its attempts to better protect freedom of expression, the drafting of the Bill risks severely undermining it.

All of these exemptions solidify the Bill's status as regulation designed with content moderation at its heart, rather than the 'systems' approach it claims to take.

Exempting from online safety measures certain categories of content and certain types of users and simultaneously requiring *additional* protection considerations for *yet other* categories of content means platforms are encouraged to focus far more on content than risky system design.

For instance, **exempting media content from a systems-based approach is not coherent**. If a platform has a truly effective and proportionate system to reduce the risk of harms to users, as required in the Bill, media content will only be negatively and disproportionately affected by that system if and where it poses a significant risk of harm to users.

The inclusion of exceptions is a tacit acknowledgement that the existing freedom of expression safeguards in the Bill are *insufficient* to protect against over-moderation beyond a platforms' safety duties - or that platform safety duties should come first when dealing with some forms of speech, but not others.

It is well-established that regulation focusing on content moderation is much more risky to freedom of expression than regulation which focuses on the wider systems and processes platforms have in place: such as the powers that users have, the way decisions about content moderation and curation are made, how new features are tested, how algorithms used are trained, deployed and assessed, and how user data and engagement is monetised.

For further explanation of the difference between a content-first approach and a systems-first approach, see here and Addendum 1.

Moreover, the drafting of the journalistic and democratic content protections lead to a paradox. The intention of these clauses is apparently to protect political and journalistic freedom of expression against over-moderation by platforms in pursuit of their safety duties.

However, the way these duties are written into the Bill is extremely vague, and are in addition to a general freedom of expression duty that should protect the freedom of expression of all users, across all topics.

Hence in practice:

- Either these 'extra' protections will have little practical consequence: in which case, the governments' promises of protecting political and journalistic speech fall back on the general freedom of expression duty which they seem to feel is inadequate to do so.

- Or it will be the case, as has been strongly suggested by the government's communications, that platforms will be expected to explicitly *not* moderate political and journalistic content to the same standard as other content.

Given the difficulties of defining what counts as 'journalistic' or 'democratic' speech, this will likely mean that **politicians and the press (including rogue outlets, such as those run by far-right extremists such as Tommy Robinson) will enjoy extra levels of freedom online regardless of the propensity of their activity to cause harm.**

Governments and civil society around the world have been desperately trying to get platforms to enforce their terms of service consistently, to protect people from serious harms. However, as we learned from whistleblower Frances Haugen, Facebook has an internal program called "cross check" that *explicitly exempts* millions of users from certain content moderation rules. **This Bill, far from challenging this status quo, would dangerously enshrine it in law.**

Specific risks of each exemption or exception for promoting abuse and disinformation without protecting freedom of speech

Media Exemption

Platforms have no duties to take any action regarding risks arising from news publisher content. However, the criteria to qualify as a news publisher which are laid out in the bill are **arbitrary and very easily met**, meaning a bad actor - foreign or domestic - could easily qualify as media and manipulate the rules. Meanwhile, other publishers, which are regulated to high standards, will not meet the terms of the exemption.

For example, the definition requires publishers to publish content which "(i) is created by different persons, and (ii) is subject to editorial control". Yet it is possible for an individual to run a hyperlocal blog, be regulated by the UK's only independent press regulator IMPRESS, and yet - being written by a single person - miss out on the exemption. Such publishers, regulated by IMPRESS, are bound by higher standards than most national newspapers - but they would not benefit from the exemption.

The media exemption in action: Research showed that Russia Today, which is backed by the Kremlin, would qualify for a media exemption. Any entity could set up a qualifying entity and claim their content should be excluded from a platforms' measures to tackle risk. The UK could quickly become the world's disinformation hub with media standards set this low because so-called media outlets, including, for example, Chinese government propaganda sites or extremists, could simply register an address in the UK, meet the prescribed requirements and spread disinformation or other harmful content which in other jurisdictions would face moderation.

Misogyny and disinformation targeting women MPs could be within the scope of the Bill as a priority harm and platforms expected to say in their terms and conditions how they would respond to the risk. However, stories such as the Mail on Sunday's piece about Angela Rayner alleging she was trying to 'distract' the Prime Minister by crossing her legs, which when shared and amplified would be outside the scope of any platform intervention. It raises a key concern as to why misogynist abuse is acceptable when published by a newspaper and circulated online but not when it is directly published by a person online. The harm caused is no different.

Moreover, a platform would be expected to have a system in place to deal with the harm arising if an ordinary user created a post that compares migrants to “cockroaches”. This post would rightly be likely taken down or demoted, according to a platforms’ terms of service, on the grounds of hate and potential incitement to violence. However, the platform would not be expected - or explicitly not allowed - to take the same action against the same claim shared on their service from a news publisher: as it was in 2015, posted by a columnist sharing a column they had published in a recognised news publisher to millions of people, despite the fact that this would carry a much greater risk of harm than if posted by an anonymous user with 10 followers. This creates a clear two-tiered system based on the publisher of the content and not on the risk of causing harm at scale.

There is also an issue of consistency: the bill requires broadcast media to be independently regulated (by Ofcom) in order to benefit from the exemption, but makes no parallel requirement of non-broadcast media (print & online).

The outlook worsens: as currently drafted, the media exemption means that platforms are not *required* to say how they will address harms arising from media content, although they may if they choose (e.g. to ensure their own policies are applied consistently).

However, recent announcements indicate the Government is considering strengthening this to insist platforms *may not* take any action to reduce the risks arising from media content on their services. This is the form of exemption which was condemned by human rights and democracy experts, journalists, and fact-checkers across Europe when it was proposed in the DSA, and which was rightly abandoned.

The European Parliament even rejected weaker forms of a media exemption such as special early redress for "media" because the nature of disinformation and computational propaganda is that it can go viral and wreak havoc in minutes or hours. The Online Safety Bill should act as a circuit-breaker for disinformation not an enabler.

Journalistic content

This covers ‘news publisher content or regulated content, generated for the purposes of journalism, and which is ‘UK-linked’. This includes, but is not limited to, content generated by news publishers, freelance journalists and citizen journalists.’

Firstly, the journalistic content exception moves towards creating a two-tier system without appropriate protections: It is suggesting that content which is deemed journalistic, regardless of what it is, whether or not it is accurate, could cause harm, or is relevant to a debate - will benefit from additional protections.

In practice, this is likely to mean that content shared by journalists is additionally protected, as demonstrating that content shared by journalists is for the purposes of journalism is clear.

Under Article 10 of the ECHR, extra protections afforded to public watchdogs, such as the press, in recognition of their role in protecting and promoting the public interest, should be a *conditional* protection: *“The increased protection afforded to “public watchdogs” and particularly*

the press under Article 10 is subject to the condition that they comply with the duties and responsibilities connected with the function of journalist, and the consequent obligation of "responsible journalism".

No such obligation or standard exists for those claiming their content is 'for the purposes of journalism' and availing themselves of an expedited complaints procedure under the Online Safety Bill.

The exception in action: A HOPE not hate report found that it is a common tactic of the far right to claim to be 'citizen journalists' in order to give their activity legitimacy: for instance, in 2020, they found that 'a handful of figures have spent large amounts of time filming the arrival of boats and various locations used to house arriving migrants, such as hotels. Their videos, which have sometimes included chasing migrants with cameras, have quickly spread across far-right social media platforms and whipped anti-immigrant activists into a peak of anger. Each new video seeks to confirm the far right's existing belief that Britain is currently under attack.'

Far from challenging this, the journalistic protections as written would enshrine in law that users of this sort (who claim their extremism is journalism) should have at minimum an expedited appeal process above and beyond other ordinary users, and likely a higher threshold for their content being taken down or demoted regardless of its likelihood to cause harm at scale.

Content of democratic importance

This covers content 'which is, or appears to be, specifically intended to contribute to democratic political debate in the United Kingdom or in any part or area of the United Kingdom. Examples of such content would be content promoting or opposing government policy and content promoting or opposing a political party.'

Secondly, exceptions for content of "democratic importance" mean that if a topic is related to democratic political debate then it is meant to be given greater protections: a category which is either so wide as to be meaningless, or (as suggested by the explanatory notes) privileges speech of those discussing government policy and political parties - likely to be disproportionately politicians themselves.

The government has made clear that preserving freedom of expression is the most important duty, meaning that other protections outlined in the bill, like safety from abuse, are subjugated if the topic is of "democratic importance." This *also* applies with regard to how platforms treat content which is harmful to children.

For example, the government is currently proposing a ban on conversion therapy that excludes transgender people from its protections. This is a subject of intense political debate across parties, meaning that a hate campaign targeting trans children and arguing that they should be subject to conversion therapy - which a UN expert has argued can amount to 'torture' - would be given special considerations.

The bill will lead to platforms effectively making decisions based on who a user is rather than the potential for harm, particularly at scale, blatantly privileging certain users over others and creating a two-tiered Internet. This is unequal and dangerous.

Harm is harm wherever it appears online and mitigation measures to remove or slow the spread of such content should be universally applied based on its likelihood to cause harm. Content online does not somehow become less harmful depending on what user posted it or whether a topic is deemed currently relevant to politics.

If the government judges that platforms' implementation of their safety duties are likely to be such that accurate news reporting is put at risk, and that political debate is stifled, these issues should be dealt with in the structure of the safety duties rather than exempting some groups from having their content affected.

This can be done by focusing on regulating the systems platforms have in place for reducing risks to their users, rather than on regulating what platforms do with different types of content. Further explanation of how this can be achieved can be found in Addendum 1.

Exclusion of paid ads

Moreover, paid ads are almost entirely out of scope, meaning users and entities can pay to evade the rules. **The content is no less harmful because it has been paid for.** In many cases, the opposite is true, *and* the platforms are directly profiting. Adverts spreading vaccine disinformation, for instance, would be excluded from what platforms would be expected to mitigate the harm of.

Far from reducing the risk of over moderation, these exemptions, exceptions and exclusions create perverse incentives for platforms: to over moderate the free expression of normal people who are more likely to fall outside these exemptions and are apparently, less deserving of freedom of expression protections, while allowing potentially hateful and abusive content to proliferate as long as it comes from people and entities that fall into these categories.

From the point of view of combating abuse and disinformation at scale, it would be better for content that is legal but harmful to adults to be out of scope of the Bill altogether than for it to be in scope but with exemptions giving some forms of abuse and disinformation a government-sanctioned free pass.

Using a defence of free speech to permit disinformation and abuse is a tactic used by other governments, notably governments which fail to uphold democratic and human rights. Poland, for example, has proposed a 'free speech law' after a far-right party's Facebook page was removed for spreading hate speech and Covid disinformation. It is dangerous to allow our liberal institutions to be co-opted for harmful aims.

Numerous additional real-world examples of how the media exemption, democratic and journalist speech exceptions and paid ads exclusions would allow for grievous offline harm are included in Addendum 2.

Ensuring the Online Safety Bill delivers: Recommended Amendment Areas

The protections for freedom of expression should be strengthened. The media exemptions, special consideration for democratic and journalistic content should be removed, and all paid ads brought explicitly in scope. We set out how this can be achieved below.

Amendment Area 1: Strengthening freedom of expression

The Online Safety Bill should be free speech preserving and harm reducing. It can do that by strengthening the freedom of expression provisions that already exist in the bill. **Every user, whether they are an elected representative, registered party representative or candidate, a newspaper, a journalist, nurse or a pensioner, deserves equal protection to participate freely and safely online and should equally be held to the same standards.**

The existing duties in the Bill to protect rights are extremely vague and overly narrow: platforms need only have regard to the importance of protecting privacy and freedom of expression, with little priority or specificity given to how this will be ensured.

This could be achieved in the following ways:

- **Schedule 4 should be amended so that the online safety objectives for regulated user-to-user services and regulated search services both include rights protections, such as including that:**
    a) A service should be designed and operated in such a way that the human rights, as defined in the Human Rights Act, European Convention on Human Rights and UN Convention on the Rights of the Child, of users and affected persons are protected, including that journalism holds a unique and central role in democratic society, and is recognised in alignment with Article 10 of the ECHR
    b) The amendment should be that in the course of its duties, in carrying out risk assessments, serving information or enforcement notices, and developing Codes of Practice, OFCOM should be required to carry out a rights impact assessment on the systems and risks that they are assessing and the systems or technologies they are recommending (Part 7, Chapter 3).

Amendment Area 2: Strengthening Online Safety

Strengthening the free speech protections in the bill would **allow for it to be amended, removing the media exemption and the exception for content of "democratic importance" and bringing paid ads into scope.**

This would ensure a level playing field for freedom of expression and would deliver on the Bill's aims to reduce harms online.

–##--Addendums Follow–##--

Addendum 1: Examples of a systems-based approach to risk mitigation

As noted in the briefing above, if the government judges that platforms' implementation of their safety duties are likely to be such that accurate news reporting is put at risk, and that political debate is stifled, these issues should be dealt with in the structure of the safety and transparency duties rather than exempting some groups from having their content affected.

This can be done by forcing open the algorithmic blackbox and focusing on regulating the systems platforms have in place for reducing risks to their users, rather than on regulating what platforms do with different types of content. It is transparency and auditing of the human decisions, algorithmic changes and data inputs that will actually shift the balance of power towards robust journalism and publishers.

How do we regulate systems rather than content?

The systems that platforms have in place can increase or decrease risks to users: it is essential for the regulator to have robust audit powers and researchers have independent access to platform data to test and measure how these risks in fact change according to different decisions made by platforms.

- **Technical systems and processes**
    - What are algorithms designed to maximise for?
    - How are algorithms being trained, and how is their efficacy measured and tested?
    - On what basis are certain trends or certain kinds of content recommended, promoted or demoted?
    - What data is collected, how is it stored, and how is it used to serve users content or advertising?
    - What kind of activities can users engage in on a platform - how are they able to communicate or interact with each other? What are they incentivised to do: where is friction introduced in their interactions?
- **Organisational systems and processes**
    - Who is involved in making decisions about platform policies?
    - Are risk and rights assessments conducted before changes are made to a platform? How are these carried out? Are they robust?
    - Are people involved in content moderation given appropriate training and psychological support? Do they have the relevant expertise in recognising different kinds of harms? How many languages do they speak?

****** Continued on next page ******

For example:

Risk of harm to user	Content approach: the regulator might ask	Systems approach: the regulator might ask
Exposure to promotion of suicide	Is all content promoting suicide taken down?	If users search for, post, share or are exposed to content promoting suicide, is there a system through which they can be directed to/access emergency and longer-term support? What does your recommender algorithm serve a user who has searched for suicide content more than once?
Exposure to vaccine disinformation	Is vaccine disinformation content demoted?	How do platforms identify when vaccine disinformation is reaching wide audiences and are mitigations (such as promoting authoritative information or fact-checking vaccine content) able to be quickly put in place?
Subjected to racist pile-on harassment	Can users report content which is harassing them?	What functionalities or design choices encourage or incentivise pile-on harassment - do pile-ons feed into 'trends', can they be easily monetised?

Moreover, rather than carve-outs and exemptions, the govt should be selling transparency as the way to tackle arbitrary harms *to publishers.*

Example 1: Radicalising Audiences
Research that analyzed over 2 million recommendations and 72 million comments on YouTube in 2019 revealed that its recommendation system steers viewers towards politically extreme content. News audiences on the centre right are being drawn further and further away from facts and rigorous journalism. YouTube also announced in 2019 that it was expanding an experimental tweak to its algorithm to stop the amplification of conspiracy theories in the UK

market. That clearly failed during Covid-19 but there is absolutely no way for journalists, MPs or civil society to measure what impact that has had.

Example 2: Non-transparent downranking

In 2017, Mark Zuckerberg instructed his engineers and data scientists to design algorithmic "ranking changes." But under pressure from a Republican administration it was made clear to engineers, "we can't do a ranking change that would hurt Breitbart." Engineers made the necessary tweaks which showed less impact on conservative sites and more harm to progressive leaning ones - one of which was Mother Jones. Mother Jones saw an "enormous decline in traffic from Facebook and a consequent decline in revenue." The full story came to light almost three years after algorithm tweaks that no one knew were being weaponized against certain outlets.

Addendum 2: Global examples of how exemptions, exceptions and exclusions will cause harm

Around the world, it is often political leaders, news publishers and paid advertisers – those who are most likely to be protected under the Bill – that can cause the most harm on social media due to their disproportionate influence and reach. As drafted, the bill exempts, excludes or allows exceptions for what extensive research around the world has shown are the most common sources of harmful content online. **This harms, rather than helps, a free and independent press.**

Taking action against abuse and disinformation is not antithetical to journalism - it *supports* journalism. Maria Ressa, Nobel Peace Prize winner and journalist, who has been the subject of horrendous online abuse, threats and disinformation campaigns, has called for 'shifting social priorities to rebuild journalism for the 21st century while regulating and outlawing the surveillance economics that profit from hate and lies.' Speaking specifically against the DSA backdoor exemptions, she tweeted that: It can only protect media freedom if it is able to cut off disinformation upstream. Creating special media redressal mechanism may sound good but will ENABLE industrial scale disinformation'. 'Especially given the low standards outlined in this bill for content producers to qualify as news publishers, exempting journalists simply allows the degradation of journalism in the 21st century to continue.

The Potential Impact of the Media Exemption

Allowing any media content that meets the bill's lenient criteria to stay up or be fast-tracked for appeal if removed enables publishers to reach wide audiences with hateful or false claims with minimal consequences. As the EU Disinformation Lab noted in a letter to IMCO Committee members, "it is virtually impossible to define who or what is a legitimate 'press publication' in the online environment". **For example:**

- Figures such as Alex Jones and Milo Yianopolis, who have been repeatedly suspended and banned from social media sites for hate and abuse, could easily be considered media figures, producing content 'for the purposes of journalism'.
- The German magazine *Compact,* categorised by German intelligence services as "demonstrably extremist" with "clear far-right aspirations", is currently available in the UK and active on Twitter. While Facebook removed *Compact* for breaching its community standards, the magazine would likely qualify for a media exemption under this legislation and would therefore be much more difficult for Facebook to moderate or remove.
- In India, a 15-year global disinformation campaign targeting the UN and the EU and aimed at "discrediting Pakistan internationally" utilised at least 750 fake local media outlets to launder pro-India propaganda as authentic content. The low bar for qualification as news publisher content means that fake outlets operating in

coordinated propaganda campaigns like this one would likely be protected under this Bill.

- Another significant concern is the rise of "content farms", which game search engines to widely disseminate low-quality, mass produced content under the guise of journalism. In the US, Facebook took action against a publisher called *Peace Data,* which lacked any editorial supervision and failed to "uphold the tenets of respectable journalism", for widely disseminating Russian propaganda as part of a potential coordinated campaign. In Taiwan, content farms based in Malaysia widely disseminated pro-China propaganda in the run-up to elections, with floods of similarly-worded "news" articles popping up simultaneously around the nation and spreading rapidly due to Search Engine Optimization (SEO). Content farming is an industry specialising in the mass production of "journalistic content" optimised for algorithmic amplification, with no regard for truth and often with nefarious intent. This Bill's exemption for journalistic content has no regard for the reality of global coordinated disinformation campaigns and the role that journalistic content can play in them.
- News media content can lead to online radicalisation that has real-world consequences. Tucker Carlson of Fox News, a recognised news publisher in the United States, has endorsed a dangerous white supremacist conspiracy theory called the "great replacement", in which ethnic minorities are supposedly replacing white Europeans in order to further a left-wing political agenda. These theories are also promoted by smaller-scale "citizen journalist" commentators online. This theory inspired a mass-shooting at a grocery store in Buffalo, New York on May 14th, with the shooter acknowledging he had absorbed such content from Carlson and other journalists online.

The Potential Impact of the Democratic Importance Exception

Political leaders, whose speech is inherently democratically important, have large audiences and thus far more power to spread harmful disinformation and dangerous speech than regular users. Around the world, it's been shown that online political speech can have tangible and harmful outcomes in the real world. **For example:**

- When Donald Trump and some of his allies were de-platformed from Facebook, misinformation, particularly related to the outcome of the 2020 election, fell 73%. This wasn't due only to Trump ceasing to tweet harmful content, but due to his followers also disengaging in their spreading of harmful content, demonstrating that the amplification associated with well-known political figures exacerbates harm in ways that generic user-generated content does not. Similarly, political leaders in Brazil, Hungary, India and other countries have all been reported to engage in political disinformation campaigns, utilising the spread of disinformation as a "political strategy". The risk posed from such high-profile content cannot be neglected by a Bill aimed at making the digital environment safer.
- More than causing "offence", dangerous political speech can also incite violence, "increasing the risk that audiences condone or participate in violence against members of another group". Dehumanising speech or speech directly calling for violence against certain groups coming from an influential speaker is clearly documented to lead to online and offline violence. Donald Trump's tweets about the "Chinese Virus" at the beginning of the COVID-19 pandemic was found to have fueled anti-Asian hate on

Twitter which then spread into the real world. After Trump's derogatory tweet about progressive politicians in the US, online threats and abuse against those he named doubled in a 3-month period. A 2019 study found a correlation between Trump's tweets about Muslims, Twitter usage rates, and the incidence of anti-Muslim hate crimes in the US. Most notably, Donald Trump's tweets have been directly linked to the violent insurrection that took place on January 6th, 2021, threatening the lives of elected officials and the sanctity of democracy itself.

- In Brazil, a survey conducted by Gênero e Número tracked violence against LGBTQ people during and after the 2018 presidential campaign. It found that over 50 percent of respondents suffered from some form of violence due to their sexual orientation. At least 92 percent claimed that such violence increased following Bolsonaro's election.
- Facebook and other platforms are already reluctant to remove severely harmful content when it comes from political leaders, calling into question Bill's intention to exclude political content when it is already severely harmful and resilient to moderation.
- In India, T. Raja Singh, a member of the Bharatiya Janata Party (BJP), called for the massacre of Rohingya Muslim refugees online, threatened to demolish mosques, and labelled Indian Muslims traitors. Facebook hesitated to ban him due to threats from the BJP about Facebook's "business prospects" in India. While he was ultimately banned in late 2020, his status as a politician enabled him to call explicitly for violence and threaten the lives of marginalised groups in India that were left up and disseminated to his supporters and their networks.
- János Lázár, Hungarian Chief of Staff to the Prime Minister, engaged in hate speech online, posting a video of him saying that letting Muslim immigrants into Hungary will result in "crime, poverty, dirt, and impossible conditions in our cities". Facebook reversed a decision to remove the video following accusations of censorship, arguing that the video was "important to the public interest".
- A number of UK political parties and former candidates have been removed from mainstream social media platforms for consistently breaking terms of service on the platforms. For example, Stephen Yaxley-Lennon, commonly known as Tommy Robinson ran in the 2019 European Elections. Britain First, a registered UK political party, who ran candidates as recently as the 2022 local elections were removed from Twitter in December 2017 and Facebook in March 2018. This has significantly curtailed their ability to spread harmful content to mass audiences but this exemption could allow them to argue for their replatforming.
- The white supremacist "great replacement theory" has been promoted by representatives of the US government, validating a hateful outlook that led to a mass-shooting in Buffalo, New York on May 14th. This speech would be protected by the "speech of democratic importance" exemption, yet undeniably caused harm in the real world that can be traced directly back to online dissemination by news media and political actors.

This Bill fails to acknowledge that political speech is already severely under-regulated and is causing real world and online harm and violence around the world under the status quo. Protecting people online involves tackling content from all types of people, whether regular users or high office holders.

The Potential Impact of the (near-total) Paid Advertising Exclusion:

The near-total exclusion of paid advertising from the Bill is a puzzling failure to recognise a fundamental source of harm online - the surveillance-based business model of large tech platforms that rely on data extraction and user attention by any means - to sell ads. (in a late concession before the publication of the Bill, the Government announced the inclusion of a new duty in the bill to bring "fraudulent paid-for adverts on social media and search engines into scope, whether they are controlled by the platform itself or an advertising intermediary").Paid advertising is widely disseminated and often poorly fact checked, allowing it to play a significant role in spreading disinformation, inciting violence, and undermining the integrity of information environments. Advertising made up almost 98% of global revenue for Facebook in 2020 and nearly 83% for Google in 2022. Targeted ads algorithmically target users en masse with content that has been demonstrably linked to online and real world harm. Excluding paid ads (except those that are deemed "fraudulent" - eg scams - in a financial sense) from the Bill's regulatory framework would lead the legislation to neglect significant harm online.

Paid ads have been shown to facilitate the spread of disinformation, harming children and even playing a role in coordinated disinformation campaigns. **For example:**

- Facebook has approved ads that contain COVID-19 disinformation, disinformation about the validity of elections, and political ads containing scams and malware. Google and YouTube have similar problems.
- On climate change and environmental issues, polluting companies and right-wing organisations spend millions on advertising designed to disinform. These ads promote false narratives about the environment, claiming that dangerous pollutants are clean and necessary, attacking environmental regulations as a threat to the economy, insinuating that politicians are lying about climate change, and positing that climate change is just weather. Ads like these are commonplace and seen by millions of people.
- Research from Australia shows Facebook approved adverts targeting teenagers interested in gambling, smoking and extreme weight loss. Despite Facebook supposedly forbidding advertising alcohol to under-18s, the researchers were able to get cocktail recipe ads approved for targeting to 52,000 teenagers who Facebook identified as being interested in alcohol.
- They have also been shown to promote hate speech, violence, and disinformation about human rights abuses. In Myanmar, Facebook approved paid advertising that incited violence and genocide against Rohingya Muslims. A Chinese state-controlled tabloid, *Global Times,* posted sponsored videos purporting to show that Uighur Muslims were learning "vocational skills" in internment camps which Facebook did not remove. Instead of simply refusing paid ads from Chinese state-controlled media, Facebook chose to passively rely on outside experts to flag problematic posts, which it may or may not then remove, at a pace that may or may not be quick enough to avert harm. In effect, Facebook was enabling China to use the platform to cover up widespread human rights abuses and violence in Xinjiang.
- Anti-abortion ads in the United States directly target women with misleading information, including that abortion was linked to infertility and breast cancer.
- Politicians in the United States used paid ads to promote the great replacement theory, a racist dog whistle that suggests white people are being "replaced" by minorities. This incited a mass-shooting in Buffalo, New York on May 14th. These political ads would be out of scope for this bill, yet contributed directly to real-world violence.