IEA Briefing Paper

AN UNSAFE BILL

How the Online Safety Bill threatens free speech, innovation and privacy

By Matthew Lesh,
Victoria Hewson

June 2022

iea

An Unsafe Bill

How the Online Safety Bill threatens free speech, innovation and privacy

Matthew Lesh, Head of Public Policy, Institute of Economic Affairs,
Victoria Hewson, Head of Regulatory Affairs, Institute of Economic Affairs
Institute of Economic Affairs
27 June 2022

Summary

- The Online Safety Bill establishes a new regulatory regime for digital platforms intended to improve online safety.
- The Bill raises significant issues for freedom of expression, privacy and innovation.
- There is a lack of evidence to justify the legislation, with respect to both the alleged prevalence of what the Bill treats as 'harm' and the link between the proposed measures and the desired objectives.

Freedom of expression

- The duties in the Bill, in respect of illegal content and legal content that is harmful to adults, combined with the threat of large fines and criminal liability, risks platforms using automated tools in a precautionary and censorious manner.
- The Bill appears designed to discourage platforms from hosting speech that the Secretary of State considers to be harmful, even if that speech is legal. The Bill allows for the expansion of the category of 'legal but harmful' content with limited parliamentary scrutiny.
- The Secretary of State and Ofcom will have unprecedented powers to define and limit speech, with limited parliamentary or judicial oversight.
- The introduction of age assurance requirements will force search engines and social media to withhold potentially harmful information by default, making it difficult for adults to access information without logging into services, and entirely forbidding children from content even if it could be educationally valuable.
- Some small to mid-sized overseas platforms could block access for UK users to limit their regulatory costs and risks, thereby reducing British users' access to online content.
- Safeguards designed to protect free expression are comparatively weak and could backfire by requiring application in a 'consistent' manner, leading to the removal of more content.

Privacy

- The safety duties will lead platforms to profile users and monitor their content and interactions including by using technologies mandated by Ofcom.
- The inclusion of private messaging in the duties risks undermining encryption.
- The child safety duties will infringe the privacy of adult users by requiring them to verify their age, through an identity verification or age assurance process, to access content that is judged unsuitable for children.
- The user empowerment duties will further necessitate many users verifying their identities to platforms.

Innovation

- The Bill imposes byzantine requirements on businesses of all sizes. Platforms face large regulatory costs and criminal liability for violations, which could discourage investment and research and development in the United Kingdom.
- The Bill's regulatory costs will be more burdensome for start-ups and small and medium-sized businesses, which lack the resources to invest in legal and regulatory compliance and automated systems, and therefore the Bill could entrench the market position of 'Big Tech' companies.
- The likely result of the additional regulatory and cost burdens on digital businesses will be the slower and more cautious introduction of new innovative products or features, and fewer companies entering the sector. This will lead to less competition and less incentive to innovate, with resulting losses to consumer welfare.

Table of contents

Introduction

The Online Safety Bill is an enormously complicated and hard to interpret piece of legislation. It aims to make the internet safer for users and hold the tech giants to account while protecting freedom of expression.[1] The Bill imposes duties on user-to-user services (e.g., social media, web forums, online gaming, private messaging) and search engines to safeguard users.[2] In this briefing we will refer to the providers of these services jointly as 'platforms'. The proposed duties on platforms include, among other things:

- duties to undertake risk assessments for illegal content,[3] to prevent or limit access to illegal content,[4] to provide content reporting and complaints procedures,[5] to protect freedom of expression and privacy,[6] and to keep certain records;[7]
- for services likely to be accessed by children, duties to undertake children's risk assessments[8] and protect children's online safety;[9] and
- for the largest, highest risk user-to-user services, known as 'Category 1' services, duties to protect users from designated content that is harmful to adults but not illegal (defined as 'priority content that is harmful to adults', informally known as 'legal but harmful')[10], to provide user empowerment tools,[11] and to protect 'content of democratic importance' and 'journalistic content'.[12] [13]

Ofcom will be responsible for regulating platforms by developing codes of practice (to clarify and operationalise the complex provisions of the Bill[14]) and monitoring compliance with the duties. It will also have powers to issue proactive technology notices,[15] undertake investigations, refer managers for criminal sanctions, and issue fines of up to 10 per cent of global revenue for non-compliance (see Part 7 for the powers and duties of Ofcom). The Secretary of State will have the power to set Ofcom's strategic priorities and direct Ofcom to modify its codes of practice. The Secretary of State will also have the power, through secondary legislation, to set the criteria for platform categorisation (such as defining the criteria for a Category 1 service) and designate priority illegal offences[16] and priority content that is harmful to children and adults. The Bill also introduces new

[1] UK Government, 'World-first online safety laws introduced in Parliament', 17 March 2022 (https://www.gov.uk/government/news/world-first-online-safety-laws-introduced-in-parliament). This briefing analyses the Bill as originally introduced by the government into Parliament in March 2022.
[2] There are over 25,000 in-scope companies including social media sites, search engines, video sharing websites, web forums and online gaming. There are also additional duties for pornography and advertising platforms that are not discussed in this briefing. The duties do not apply to emails, SMS and MMS messages, one-to-one live aural communications, comments and reviews on provider content (for example, Amazon reviews) or 'news publisher content' (that is, articles published by a recognised news publisher). They must have links to the UK, which include being based in the UK or having British users.
[3] s. 8 for user-to-user services, s. 23 for search services
[4] s. 9 for user-to-user services, s. 24 for search services
[5] s. 17 and 18 for user-to-user services, ss. 27 and 28 for search services
[6] s. 19 for user- to-user services, s. 29 for search services
[7] s. 20 for user-to-user services, s. 30 for search services
[8] s. 10 for user-to-user services, s. 25 for search services
[9] s. 11 for user-to-user services, s. 26 for search services
[10] ss. 12 and 13
[11] s. 14
[12] ss. 15 and 16
[13] There are also additional duties for Category 2A, the largest search platforms, and Category 2B, mid-sized to large user-to-user platforms.
[14] s. 37
[15] This refers to automated monitoring technology, defined in the Bill as (1) content moderation technology including use of certain algorithms, (2) user profiling technology, or (3) behaviour identification technology (s. 184).
[16] Platforms are required to proactively remove illegal content that is designated as priority, as will be discussed further below.

harmful online communications offences that will apply to individuals, replacing Section 127 of the Communications Act.[17] [18]

This briefing (1) considers the Bill's impact on freedom of speech, privacy and innovation; and (2) raises questions about the evidence base for the proposed legislation.

This briefing does not attempt to summarise the entire Bill or outline its full scope, but rather focuses on relevant sections with respect to new duties on platforms. There are significant provisions in the Bill that this briefing does not cover in detail, such as new communications offences and duties relating to pornographic content and fraudulent advertising.

Free speech

This section describes the implications of the Bill for freedom of expression.

The risk of too much content being removed

The Bill's illegal content safety duty requires platforms to use 'proportionate systems and processes' to (a) prevent users from encountering 'priority illegal content', (b) minimise the length of time priority illegal content is present, and (c) and 'swiftly take down' other illegal material that the platform is alerted to.[19] Priority illegal content is defined as child exploitation content and terrorism content and content that 'amounts to' offences listed in a schedule (Schedule 7) that the Secretary of State can update.[20] The initial schedule includes hate crime offences under the Public Order Act 1986, harassment and stalking, drug-related offences, incitement to violence, encouraging suicide, revenge and unlawful extreme pornography fraud, money laundering, organising prostitution for gain and organised immigration offences.

The basis on which this material must be removed or minimised, or on which users must be 'prevented from encountering' it, however, makes it likely that too much – or far too much – material will be removed. Section 9(3) of the Bill sets out the statutory duties described above (systems and processes designed to prevent encounters with, minimisation and removal of illegal content). Sections 9(5) and (6) require user-to-user platforms to include provisions in their terms of service specifying how users are to be protected from 'illegal content' as it is categorised in Section 9(3), and to apply the terms 'consistently' in relation to content that the platform 'reasonably considers is illegal content or a particular kind of illegal content' (equivalent provisions apply in respect of search services). It is not clear how the terms of service are intended to relate to the statutory duties with respect to the systems and processes. In addition to causing legal uncertainty, even if not the intention of the drafting, the terms of service duty seems likely to lead to the threshold for removal and minimisation being lowered in practice to apply to content that is reasonably considered to be illegal. By any interpretation, considering the large quantity of content and the threat of large fines, platforms will have to (and in some cases expressly be required by Ofcom) use automated systems to 'protect' users and fulfil the minimisation and removal duties, and they will be incentivised to do so in a cautious and censorious manner against content that is

[17] The new communications offences are set out in Part 10 of the Bill, as discussed further later in this paper.

[18] The Bill also makes various other changes, some of which are unobjectionable, such as requiring platforms to report child exploitation material to the National Crime Agency (s. 59), formalising and placing on a statutory footing current reporting practices.

[19] Other material in-scope of the illegal content duty is defined as any criminal offence 'for which the victim or intended victim is an individual', not including economic crimes (s. 52(4)(d)).

[20] s. 176

only reasonably considered to be illegal.[21]

The obligations created by Section 9 are not simply duties to apply the same legal principles to online material as already apply offline.[22] In the case of offline offences, such as 'hate crime' offences under the Public Order Act, there is a requirement for consent from the Attorney General's Office before proceedings can be brought, and if the matter is pursued, the criminal law includes many protections for civil liberties and freedom of expression. For platforms to take action against material they 'reasonably consider' to be criminal evidently sets a much lower threshold. Even if platforms are only required to remove content that 'amounts to' a criminal offence (the definition of illegal content under Section 52(3)), this places them in the role of a judge, but without the legitimacy or contextual information necessary for the role.

Further concerns arise in connection with new offences created by the Bill. The 'harmful communications offence' created by the Bill consists of sending a message that is intended to cause at least serious distress to a likely audience.[23] [24] The 'false communications offence' is defined as sending information known to be false that causes non-trivial damage.[25] [26] Since material which does either of these things will be 'illegal content', the illegal content safety duty will require platforms to swiftly remove content that amounts to these offences when they are alerted to or become aware of its presence. This will necessitate a platform to somehow determine a poster's intent (as the new offences require intent to cause harm or spread false information). It is difficult to see how they can reasonably do that.

A likely outcome is that those who are easily offended or who are acting in bad faith will procure removals by claiming material to be intentionally false or psychologically distressing to a 'likely audience'. That places the burden on the platform to remove it or risk non-compliance with its duty, and potential fines and other sanctions from Ofcom.[27] Clearly, material is likely to be removed even if it only has the *potential* to cause distress or *could be* intentionally false. One might consider, for example, a joke using irony to mock anti-vaxxers, which includes various descriptions of material known to be false. If a platform were to receive a complaint about that, it would be difficult for it to be sure it was safe not to remove it.

The independent Regulatory Policy Committee highlighted the risk that 'individuals or groups could seek to shut down legitimate debate or criticism by influencing decisions of [social] media companies and Ofcom through misuse of user reporting functionality' and 'seek to disable the mechanism or to impose costs on service providers by spurious or excessive requests'.[28] This

[21] The explanatory memorandum discusses how the UK will no longer be maintaining the EU's prohibition on mandating general monitoring of content; the Bill also allows Ofcom to mandate the use of proactive technology to comply with illegal content duties.

[22] The claim that the Bill merely replicates the position with offline speech was made, for example by Damien Collins MP (https://committees.parliament.uk/committee/534/draft-online-safety-bill-joint-committee/news/159784/no-longer-the-land-of-the-lawless-joint-committee-reports/)

[23] s. 150

[24] The Bill states that a person commits this offence if the person sends a message that 'there was a real and substantial risk that it would cause harm to a likely audience' and 'the person intended to cause harm to a likely audience' and 'the person has no reasonable excuse for sending the message'. 'Harm' in this context means psychological harm amounting to at least serious distress.

[25] s. 151

[26] The Bill states that a person commits this offence if a person sends a message that 'the person knows to be false' and the person intended the message to cause 'non-trivial psychological or physical harm' and has 'no reasonable excuse for sending the message.

[27] The government claims that the Bill focuses on systems and processes, rather than individual pieces of content; however, in practice it is difficult to see how it will be possible for Ofcom to monitor compliance without examining platforms' responses to individual pieces of content.

[28] Regulatory Policy Committee, 'The Online Safety Bill: RPC opinion', 18 February 2022 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1061266/2022-02-18-RPC-DCMS-4347_4_-_Online_Safety_Bill.pdf).

identifies the danger precisely.

While these considerations suggest the Bill contains significant threats to free speech, it may also be ineffective in achieving its 'safety' objectives. The Independent Reviewer of Terrorism Legislation has advised that, as it stands, the Bill would be ineffective against terrorism content, and it could only be made effective with 'insupportable' assumptions that do not allow sufficiently for freedom of speech.[29] This is because almost all terrorism offences require intent on the part of the perpetrator, and have defences available, whereby particular conduct may or may not constitute an offence depending on context. As it is very unlikely that platforms will have information about users' states of mind and possible legal defences, it will be impossible for them to identify content that 'amounts to' a terrorism offence unless they make assumptions about state of mind and context. Such assumptions, which the Independent Reviewer considered to be insupportable and are not currently reflected in the drafting of the Bill, would be overinclusive of legal content. The same issue will apply to illegal content more generally.

The inclusion of 'legal but harmful' speech

The Bill requires Category 1 services, the larger user-to-user platforms, to specify in their terms of business how they will treat 'priority content that is harmful to adults'.[30] This leaves it open to them to state a policy of taking no particular action, or even of *promoting* such content. It is clear, however, that the risk assessment process (under Ofcom's codes of conduct) and the threat of Ofcom action, are designed to encourage minimisation and removal of such material.

At present, most large platforms have content moderation rules against many types of legal but harmful speech; the government clearly considers that this is inadequate, and the Bill is intended to lead to stricter enforcement of moderation against legal content. The stated intention of the Bill is to reduce online harm (not just illegal content): the Bill's impact assessment, for example, refers to the reduction in 'harmful content such as mis-information' as a benefit of the legislation, confirming the intention to control speech that the government considers legal but harmful. The Regulatory Policy Committee also raised concerns that the government has failed to consider properly the risks to freedom of expression presented by the inclusion of 'legal but harmful' speech in the law.[31]

The government intends to formally specify categories of 'priority content that is harmful to adults' in due course. Public statements indicate they are likely to include 'misinformation and disinformation', 'misogynistic abuse', abuse, harassment and material encouraging self-harm or eating disorders.[32] Supporters of including 'legal but harmful' speech in the legislation have suggested it could extend to 'Covid disinformation' and 'climate change denial'.[33] In the same vein, Shadow Department for Culture, Media and Sport (DCMS) Secretary Lucy Powell has raised concerns that the Bill as it stands would allow 'incels' and 'climate deniers' to 'slip through the net'.[34] She clearly envisages an extension of the notion of 'harmful' to cover matters of public policy debate. The current and future governments will have discretionary power to add to the 'legal but harmful' list – by claiming a certain phenomenon is 'harmful' (see Box 1) and laying a statutory instrument before Parliament. That process very rarely leads to meaningful debate. The Bill also

[29] Independent Reviewer of Terrorism Legislation, 'Missing Pieces: Terrorism Legislation and the Online Safety Bill', 20 April 2022 (https://terrorismlegislationreviewer.independent.gov.uk/missing-pieces-terrorism-legislation-and-the-online-safety-bill/).
[30] s. 13
[31] Regulatory Policy Committee, 'The Online Safety Bill: RPC opinion', 18 February 2022 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1061266/2022-02-18-RPC-DCMS-4347_4__-_Online_Safety_Bill.pdf).
[32] gov.uk, 'Online Safety Bill: factsheet', 19 April 2022 (https://www.gov.uk/government/publications/online-safety-bill-supporting-documents/online-safety-bill-factsheet).
[33] Response to the DCMS Sub-Committee on Online Harms and Disinformation, September 2021 (https://committees.parliament.uk/writtenevidence/38548/pdf/).
[34] 'Loophole in online safety laws "will let incels spread extremist views"', *The Telegraph*, 13 April 2022 (https://www.telegraph.co.uk/news/2022/04/13/loophole-online-safety-laws-will-let-incels-spread-extremist/).

requires user-to-user platforms to notify Ofcom when they become aware of non-designated content that is harmful to adults,[35] creating a further impetus for expanding the list. There is no provision, although it is required under human rights law, for a close examination of the factual justification and necessity of restrictions on free speech.[36] Consequently, there is a clear risk that there will be a never-ending expansion of the range of controlled content.

The risk assessment duty placed on the platforms in relation to legal but harmful speech requires them to take account of the impact of the speech on users with 'certain characteristics or members of a certain group' – the characteristics and groups remain undefined by the legislation. A government minister has previously suggested that the characteristics in scope are 'not restricted to those protected under the Equality Act 2010, ensuring that harmful content concerning other qualities (such as personal appearance) is also covered'.[37] This raises the risk that individuals or interest groups claiming distress will request the removal of speech with which they disagree. Platforms would again be in the position of needing to consider the risks of being found non-compliant with the law, and are therefore likely to remove more material than is strictly required by the Bill, or even than intended by the drafters.

Box 1: The definition of 'harm' in the Online Safety Bill

The Bill states that 'harm' means 'physical or psychological harm' – taking into account the harm that may arise from 'the nature of the content', 'the fact of its dissemination' and the 'manner of its dissemination'.[38] [39] This includes, the Bill states, harm arising from content that results in individuals acting in a way that harms themselves or another individual or increases the likelihood of harming another individual. This expressly includes 'where individuals act in such a way as a result of content that is related to that other individual's characteristics or membership of a group'.[40] There is no explicit protection in this definition of harm for speech that is in the public interest or for purposes of debate.

This is a broad, fluid and subjective definition of harm that could encompass many categories of speech. The Online Harms White Paper, for example, listed 24 potential harms. This included child sexual exploitation and terrorism, but also trolling, disinformation and, for children, 'excessive screen time'.

Limiting access to information

There is a risk that some platforms will respond to the Bill's duties by limiting the content that British users can access.

The Bill's child safety duties require platforms to consider users to be children by default, until they establish that their service is not likely to be accessed by children.[41] Platforms that are likely to be accessed by children (defined as being both possible for children to access and either known to have a significant number of child users or otherwise to be attractive to children) will need to verify

[35] s. 13(7), see Box 1 for a further discussion of the definition of 'harmful'.
[36] Gavin Millar QC argues that the Bill is incompatible with Article 10 of the European Convention on Human Rights (https://www.matrixlaw.co.uk/news/gavin-millar-qc-contributed-to-a-legal-analysis-of-the-impact-of-online-safety-bill-on-freedom-of-expression/).
[37] See https://committees.parliament.uk/publications/6336/documents/69560/default/; notably the Equality Act prohibits discrimination on the basis of specified protected characteristics of age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, and sexual orientation.
[38] s. 187
[39] There is a separate definition of harm in the new communications offences in Part 10.
[40] s. 187(4)(b)
[41] s. 11(14) for user-to user services and s. 26(14) for search

users' age through age verification or profiling technologies such as behavioural biometrics. Ofcom will be providing guidance on age and identity verification. Only once this verification is complete will platforms be allowed to grant access to content that may be harmful to children.[42] That means a search engine like Google, a forum like Reddit or an encyclopaedia like Wikipedia would have to withhold any information that could be psychologically harmful to children until a user's age had been verified. Such material would not only be pornography and child exploitation material but could also be, for example, news or historical reporting about war and violence, or scientific information about distressing issues like Covid-19. Under 18s are intended to be barred from such content entirely – notwithstanding the evident possibility that some of it is educationally valuable.

Some platforms, particularly smaller ones, may respond by entirely blocking access to children, to avoid building and operating a separate version of their product for children, with the associated increased regulatory risk. Non-UK-based services that do not wish to comply with the extensive regulatory duties may entirely block access to their service for British users – as over 1,000 US news websites did following the European Union's introduction of GDPR.[43] DuckDuckGo, a US search engine that emphasises privacy and refuses to gather information on or profile its users, may be unable to fulfil the duties to identify and tailor search results to users based on their age. Its only option, to avoid violating UK law, may be to block access for British users. The blocking of non-UK services could result in British users using virtual private networks (VPNs) to access them, thereby undermining the ability of the Bill to achieve its stated purpose.

The weak free speech protections

The Bill establishes a duty on platforms to 'have regard' to the importance of protecting users' freedom of expression in their safety measures and policies.[44] Although a failure to comply could result in enforcement action, including a fine, a duty to 'have regard' is significantly weaker than the safety duties, which include strict obligations to carry out the prescribed actions.

Category 1 services are required to assess the impact of their safety measures and policies on, and publicly state steps taken to protect, freedom of expression.[45] They also have a duty to protect content of democratic importance[46] and apply this 'in the same way to a wide diversity of political opinion'.[47] There is further duty to protect 'journalistic content', and establish a separate complaints procedure to allow users to object to the removal of journalistic content,[48] but only for UK publishers.[49] In practice, however, this will mean the free speech rights for content that is decided to be of democratic importance or is considered to be journalistic will have stronger protections than those of general users. But the scope of this protection is narrow. For example, content of democratic importance is defined by reference to 'democratic political debate'. This means issues that are of social or scientific importance, rather than being political in nature, are excluded. It is not clear whether this section would protect discussion of matters such as: (a) whether the Covid-19 pandemic was the result of a leak from a lab engaged in 'gain of function' research; (b) whether the increase in the number of children presenting with gender dysphoria is a result of social contagion; or (c) whether the IPCC understates, or overstates, climate-related risks. Even in relation to more

[42] Content that is harmful to children consists of priority content designated by the Secretary of State or content that 'presents a material risk of significant harm to an appreciable number of children in the United Kingdom' (s. 53).
[43] 'Report: 1/3 of top US news sites block EU users rather than comply with GDPR', Martech, 8 August 2018 (https://martech.org/report-1-3-of-top-us-news-sites-block-eu-users-rather-than-comply-with-gdpr/)
[44] This is a general duty that not only applies to their specific safety duties under the Bill but also to their general application of content reporting and complaints procedures.
[45] s. 19(5–7)
[46] Content of democratic importance is defined in the Bill as 'intended to contribute to democratic political debate in the United Kingdom or a part or area of the United Kingdom' (s. 15(6)).
[47] s. 15
[48] Journalistic content is defined in the Bill as either 'news publisher content' (that is, from a recognised UK news publisher) or other UK-linked content that is 'generated for the purposes of journalism' (s. 16(8)).
[49] s. 16

narrowly 'political' matters (such as law reform proposals), the scope of Section 15 is unclear. For instance, if someone advocates for the criminalisation of all abortion, are they engaging in 'democratic political debate', even though none of the political parties is currently taking such a position?

Overall, inclusion of free speech protections indicates an effort to compensate for the other sections of the Bill that undermine freedom of expression. This in itself is an admission that the duties imposed by the Bill do otherwise threaten free speech. But the protections appear wholly inadequate in any case.

Ofcom will have the power, through codes of conduct,[50] to decide how platforms should 'have regard' to freedom of expression and which content is of 'democratic importance' and what content could be harmful.[51] While it may be hoped that Ofcom will use this to ensure a vigorous defence of free speech, it will have the power to limit the parameters of discourse on public platforms. A regulatory agency will therefore be determining the bounds of online free speech.

The danger of requiring 'consistent' application of terms of service

Platforms will be required to apply the content policies in their terms of service in a 'consistent' manner,[52] and user-to-user platforms must inform users of their right to make a claim for breach of contract if content is removed inconsistently.[53] [54] This appears intended to minimise the risk of political or other bias in content moderation. However, it could have the unintended consequence of causing more content removal. This is because, at present, the terms of service for social media platforms typically allow platforms to moderate and remove content that fails to meet community standards or user terms – but do not require them to do so.[55] This is important for managing their liability, as otherwise, users could pursue legal action for failure to remove content. There are genuine uncertainties in content moderation, with many grey areas between content that is clearly against policy and content that could, by some people's interpretation but not by others, constitute harassment or discrimination. The platforms often use their discretion to not remove speech that some might consider contrary to their stated policies. The Bill's consistency duty will mean all violative material must be removed (even where a discretionary approach may have let it stand as being on balance not harmful). It may also lead to removal of content which is in fact not violative if it is arguable otherwise and retaining it risks making the platform appear inconsistent. Thus, the requirement to be consistent may result in the precautionary removal of benign material, in a drive to retain consistency and to avoid both contractual and regulatory liability. This is reminiscent of the issues raised by the 'fairness doctrine' in US broadcast communications law, which resulted in broadcasters opting to air discussions on fewer issues to avoid also having to broadcast an alternative perspective.[56]

[50] Platforms will be strongly incentivised to follow Ofcom codes of conduct, as doing so will be taken as evidence of compliance with the legal duties (s. 46) and in some cases compliance with the Code is *conclusive* evidence of compliance with the Act (albeit other means of compliance may also be available) (s. 45).
[51] ss. 45(2) and 45(3)
[52] s. 9(6), s. 11(6), s. 13(6) for user-to-user services and s. 24(6) and s. 26(6) for search
[53] s. 19(4)
[54] There is a requirement to apply safety duties about illegal content, protecting children, and protecting adults, duties to protect content of democratic importance and journalistic content, in a consistent manner.
[55] Twitter's Terms and Conditions state that the company 'reserve[s] the right to remove Content that violates the User Agreement,' but does not require it to do so (https://twitter.com/en/tos#intlTerms).
[56] See The Heritage Foundation, 'Why The Fairness Doctrine Is Anything But Fair', 29 October 1993 (https://www.heritage.org/government-regulation/report/why-the-fairness-doctrine-anything-fair).

Privacy and data security

This section describes the implications of the Bill for user privacy.

Proactive monitoring of user speech and the risk to encryption

Platforms are currently not liable for illegal content until after they are aware of its presence on their service.[57] The Bill's illegal content duties, however, will require platforms to use 'systems and processes' to ensure users *do not encounter* priority illegal content. In practice, this will necessitate general monitoring of user content using automated tools. General monitoring gives rise to serious user privacy concerns – as has been recognised for decades by the prohibition on states requiring general monitoring under the EU's e-Commerce Directive.[58] This is because it creates a mandate to actively assess every element of user speech, including in private forums, in furtherance of the objectives of the state. Such monitoring of communications in other media would be subject to strict requirements for warrants and judicial oversight, and must be based on legal grounds (see, for example, the Investigatory Powers Act 2016). Requiring private bodies to monitor and censor user content subverts these protections.

Additionally, the risk assessments for illegal content, and harm to children and adults, will require profiling of the user base by platforms in order to assess their risk profile, as either potential perpetrators or victims. Under the UK GDPR, using personal data for profiling individuals and automated decision-making that has legal or other significant effects is considered to present a high risk to their privacy and to be susceptible to errors, bias and discrimination. To the extent that individual level data is used (and in respect of some duties, it is hard to see how this could be avoided) at a minimum, a detailed assessment of impacts on data protection will be required for such processing to be lawful and specific amendments to the Data Protection Act 2018 may be required to ensure that platforms have legal grounds for such actions.

Ofcom will have the power to issue 'proactive technology' requirements in respect of illegal content, children's safety and fraudulent advertising. These could include requirements that platforms use automated systems to monitor and remove non-private content.[59] Ofcom will also be able to mandate the use of proactive technologies in private messaging in relation to child exploitation material.[60] More broadly, it is hard to see how the safety duties on private messaging services could be enforced without general monitoring of private communications. This monitoring could undermine end-to-end encryption on platforms such as WhatsApp, iMessage, Telegram and Signal. The undermining of encryption would put at risk the ability of dissidents and whistleblowers to communicate securely, opening individuals up to blackmail and putting at risk our security and privacy, both from hackers and from government snooping.

The use of automated systems to scan and categorise content is the only possible way for larger platforms to moderate their services and meet the requirements for 'systems and processes' under the Bill. While larger platforms already use automated tools and artificial intelligence (AI) for this purpose, it is widely acknowledged that, while improving in their accuracy, such solutions are imperfect even for the moderation currently carried out, and even when backed by human review.[61]

---

[57] Article 14 of the EU's e-Commerce directive, retained in UK law, protects providers from liability for illegal content unless they have knowledge or awareness of it and if they, upon said knowledge, act 'expeditiously to remove or to disable access to the information'. Some platforms use automated systems to remove illegal content, however they are not required by law to do so.
[58] Which has been carried into UK law, but will be repealed by this legislation.
[59] s. 116
[60] s. 103
[61] See, for example, 'The Limitations of Automated Tools in Content Moderation', New America, (https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/the-limitations-of-automated-tools-in-content-

As noted above, given the nuances of factual and legal analysis required to establish whether a crime has taken place, it seems unlikely that current or near-future systems will be capable of making reliable determinations as to legality, or harm. Given the intentions of and incentives under the Bill, over removal seems inevitable. Use of AI and machine learning is also controversial as the operation of such systems is not always explainable, which can lead to opacity, and could cause tensions with other duties of consistency and regard for free expression.

The need to undertake age verification

The Bill requires platforms that are likely to be accessed by children[62] to prevent anyone under the age of 18 from accessing material that could be harmful.[63] All platforms could be likely to be accessed by children unless they have systems or processes that mean children are not normally able to access the service or a relevant part of it. This can be achieved 'for example, by using age verification or another means of age assurance'.[64] This would, in practice, mean users would have to enter a passport, driver's licence or credit card number, or platforms would use profiling technology like behavioural biometrics, to confirm a user is an adult and therefore permitted to access child-inappropriate search results on Google and grown-up discussions on Mumsnet or to use YouTube beyond YouTube Kids, for example. The gathering of this additional information on users across various platforms raises significant user privacy and data security issues. It could mean users regularly being required to hand over personal identifying information to access websites, and raising risks that this information could be hacked and released.

Separately, the Bill also mandates age verification for pornography[65] – that means blocking access to pornography unless a user links their identity (drivers' licence, passport, credit card, etc.) to their adult viewing habits. This is private information that, if hacked and released, could be used to embarrass or blackmail millions of people – a possibility highlighted by how the dating website Ashley Madison data breach resulted in extortion attempts and suicides.[66]

Platforms will have a duty to 'have regard' to the importance of protecting users from breaches of privacy law. The gathering of additional user data, however, will likely be permissible under privacy law as necessary to fulfil legal obligations. The need to 'have regard' does not, in any case, add anything meaningful to the legal obligations that platforms are already under in relation to privacy and data protection.

The introduction of digital identity

The user empowerment duty requires Category 1 platforms to allow users to choose to only interact with verified users.[67] This means providing functionality for users to confirm their identity, through a driver's licence, passport or credit card, or by a recognised digital ID. This system could effectively mandate ID for the unrestricted use of online services.

In practical terms, however, it is difficult to see how this could work on global platforms, where the ability to interact with users around the world is highly valued. Users in other jurisdictions will not

---

moderation) or Gillespie, T. (2020) 'Content moderation, AI, and the question of scale, *Big Data & Society* (https://journals.sagepub.com/doi/full/10.1177/2053951720943234).

[62] 'Likely to be accessed by children' is defined as possible for children to access the service or a part of it and either in fact either used by a significant number of children or attractive to children as determined by a children's access assessment (s. 31(3))

[63] s. 11 for user-to-user services and s. 26 for search

[64] s. 11(3)(a14) for user-to-user services and s. 26(14) for search

[65] See Part 5.

[66] See discussion in 'Extortionists Target Ashley Madison Users', Krebson Security, 21 August 2015 (https://krebsonsecurity.com/2015/08/extortionists-target-ashley-madison-users/) and 'Ashley Madison: "Suicides" over website attack', BBC News, 24 August 2015 (https://www.bbc.co.uk/news/technology-34044506).

[67] s. 14

be required to identify themselves to the UK standards adopted by Ofcom. UK users who opt to only interact with verified users could therefore be facing a very reduced, UK-only version of user-to-user services online. This calls into question the usefulness and likely uptake of the facility.

Innovation

This section describes the implications of the Bill for start-ups, competition and innovation.

Time consuming and costly duties

The impact assessment estimates that more than 25,000 companies have duties imposed on them by the Bill. This includes 'Big Tech' brand names such as Twitter, Facebook, YouTube, Google, smaller networks such as Mumsnet, Change.org, and Wikipedia, trade union forums and microcommunities, such as whisky tasting forum Dram.io,[68] and private messaging services like iMessage, WhatsApp, Signal and Telegram.

The Bill imposes duties on all platforms to conduct complex risk assessments to determine whether their services are likely to be accessed by children, or contain illegal content, content that is harmful to children and (for Category 1 services) content that is harmful to adults. They will have to assess their user base, the risk of users encountering each type of illegal content, taking into account algorithms, speed of dissemination, functionality and design of the platform, and the nature and severity and the risk of harm presented to individuals. They will also have to operate complaints processes both for claims about not removing a piece of content and for removing too much content,[69] and multiple other aspects of the Bill's requirements that are in tension, if not outright conflict, with each other. There will be more specific requirements following Ofcom's development of a register and guidance about risk assessments. The Bill then requires platforms to develop proportionate systems and processes to mitigate and manage the risks of harm while having regard to freedom of expression and privacy.

The Bill mandates further risk assessments every time the platform makes a significant change to the design or operation of the platform. Ofcom is required to carry out sector level risk assessments and platforms must take into account of this, and update their own risk assessment whenever Ofcom makes a change to its assessment. Furthermore, the regulated companies will be required to undertake recordkeeping and reviews, publish transparency reports,[70] report to Ofcom, all while having regard to freedom of expression and privacy. Many companies may also need legal advice merely to determine whether the Bill regulates them, and then to remain alert as legal understanding changes over time.

These burdens will be a significant disincentive to innovate for the over 25,000 companies that are in the scope of the Bill, especially for the smaller, challenger firms for whom the legal and compliance costs will be most onerous. This could replicate the unintended consequences of GDPR, which a recent study concluded led to an 8 per cent reduction in profits for smaller firms while having no effect on the profitability of large technology companies.[71]

The impact assessment estimates that implementing the Bill will cost businesses £2.5 billion over the first ten years.[72] This figure significantly underestimates the direct costs and does not even attempt to assess the potential costs to innovation, competition, or international trade. The impact

[68] See the founder of Dram.io's view of the impact of the Bill at https://derickrethans.nl/online-safety-bill.html.
[69] s. 18 for user-to-user services and s. 28 for search
[70] ss. 64 and 65
[71] 'How data privacy regulation shaped firm performance globally, VoxEU, 10 March 2022 (https://voxeu.org/article/how-data-privacy-regulation-shaped-firm-performance-globally).
[72] See page 2 of *The Online Safety Bill: Impact Assessment*, 31 January 2022 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1061265/Online_Safety_Bill_impact_assessment.pdf)

assessment asserts that it will cost businesses, on average, £700 over ten years to read and understand the regulations, for example.[73] However, this would not realistically cover the fees of a specialist law firm for two hours, let alone the internal staff time costs. The impact assessment specifically assumes staff will only require 30 minutes to familiarise themselves with the requirements of the 255-page legislation and 90 minutes to read, assess and change terms and conditions in response to the requirements. Legal advice is estimated to cost £39.23 per hour,[74] an order of magnitude less than the fees of hundreds of pounds per hour typically charged by lawyers in this field.

The impact of criminal sanctions

Regulated companies will be required to identify a senior manager, who will be responsible for compliance, by name.[75] This individual becomes criminally liable for failure to comply with information requests from Ofcom, interfering with information requested by Ofcom, blocking power of entry by Ofcom, or providing false information in an interview with Ofcom.[76] The risks of criminality (which could ensue from carrying legal but harmful content) and court action will lead to risk averse management and could discourage companies from operating in the UK. This would mean fewer digital services accessible to British users, lessening competition and the associated benefits for innovation.

The expansive role of Ofcom and Secretary of State

A further regulatory burden is the uncertainty presented by the powers held by Ofcom and the Secretary of State. Ofcom will have the ability to set standards, undertake oversight and enforcement, impose large fines and apply to a court to block access to the platform for UK users.[77] The Secretary of State will be able to direct Ofcom to change codes of practice 'for reasons of public policy'[78] and will set Ofcom's priorities.[79] There is a significant risk that, in response to popular attention, there will be an ever-changing list of priorities and focus. Any regulated company could be investigated and, if found in breach of the rules, subsequently punished by the regulator. The scale of the content and number of companies being regulated means that, as with data protection enforcement by the Information Commissioner, Ofcom will inevitably be required to be selective in the platforms and types of harm that it takes action against. This could lead to legal uncertainty and political or other outside influence in enforcement. In practice, this level of discretion empowers Ofcom officials, whose decisions can only be challenged on procedural rather than substantive matters.[80]

Service providers overseas will have to familiarise themselves with UK criminal laws, as well as the regulatory requirements of the Bill. The impact assessment dismisses the possibility of providers exiting the UK market on the basis that the UK is a substantial international market. Nevertheless, this must be a material risk. When the GDPR came into force, many overseas websites blocked access from the EU and in some respects, this Bill would be more onerous and harder for non-UK platforms to comply with, other than for the tech giants that have bases in the UK and can take extensive legal advice.

---

[73] The impact assessment (ibid: 2) estimates that it will cost between £9.6 million and £17.5 million over ten years for businesses to read and understand the regulations, with an estimated 25,100 platforms in-scope (ibid: 18). The upper estimate (£17.5 million / 25,100 platforms) equals £697 per platform.
[74] See point 154, on page 37, of the impact assessment.
[75] s. 87
[76] ss. 92–96
[77] ss. 110–129
[78] s. 40
[79] s. 143
[80] The Bill only provides for judicial review challenges to Ofcom's decisions, which are restricted to narrow procedural or rationality grounds, giving a high level of deference to the regulator.

Evidence for the existence of a problem or the effectiveness of the proposed solutions

The Bill is expressly intended to reduce online harm. However, the government has presented limited evidence about the extent of online harm and little reason to believe that the Bill's provisions will address the harms that they claim exist.

The extent of online harm

The impact assessment, which outlines the government's justifications for the Bill, states that the quantitative assessment to demonstrate the scale of the problem relies on 'a number of uncertain assumptions, proxies and experimental data'.[81] For example, the impact assessment refers extensively to polling carried out by Ofcom to support its position that illegal and harmful content is widespread online – despite self-reported perceptions of harm being notoriously unreliable.[82] Data on crimes with an 'online element' is acknowledged in the impact assessment to be subjective and inconsistently recorded.

The Joint Committee on the draft Online Safety Bill noted[83] that 'academic research which has systematically examined the prevalence of online content that creates a risk of harm consistently finds that its prevalence is low. Abusive content, for example, made up less than 1 per cent of overall content online according to a 2019 study'.[84] The Committee went on to note that reported experience of such content is at a much higher rate, and suggests that this could be because harmful content is amplified more widely. The difference between self-reported harm and objectively measured levels of harm is a significant issue though, and one that deserved further research and consideration by the government before embarking on legislation.

The government assumes that problems like racism and bullying, or 'misinformation' and child exploitation, can be blamed on platforms that provide a conduit for human behaviour.[85] However, these issues are far longer running and deeper. For example, the impact assessment highlights Office for National Statistics data saying that around 20 per cent of children have experienced online bullying, without mentioning that the Department for Education found around 41 per cent of children reported being bullied in 2005, before the proliferation of social media. This suggests bullying may be a bigger problem offline than online and is a societal issue that is not amenable to a technical regulatory fix.[86] The impact assessment also failed to consider any alternative policy options.

---

[81] 'Experimental data', as described by the Office for National Statistics, refers to a series of statistics that is in the testing phase and not yet fully developed.
[82] See, for example, Brenner, P.S. and DeLamater (2018) 'Lies, Damned Lies, and Survey Self-Reports? Identity as a Cause of Measurement Bias', Social Psychology Quarterly (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5639921/).
[83] 'Report – Draft Online Safety Bill', Draft Online Safety Bill (Joint Committee), 14 December 2021 (https://committees.parliament.uk/committee/534/draft-online-safety-bill-joint-committee/publications/).
[84] The original source is The Alan Turing Institute (2019) *How Much Online Abuse is There? A systematic review of evidence for the UK: Policy Briefing – Summary* (https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_summary_24.11.2019_-_formatted_0.pdf)
[85] For a discussion about the ongoing debate about the extent to which social media is responsible for societal ills, see Lewis-Kraus, G. (2022) 'How Harmful *Is* Social Media?', *The New Yorker* (http://newyorker.com/culture/annals-of-inquiry/we-know-less-about-social-media-than-we-think).
[86] 'Longitudinal Study of Young People in England cohort 2: health and wellbeing at wave 2', Department for Education, July 2016 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/599871/LSYPE2_w2-research_report.pdf). A separate non-government study indicates around half of children experienced bullying in 2017: 'The Annual Bullying Survey', Ditch the Label, 2017 (https://www.ditchthelabel.org/research-papers/the-annual-bullying-survey-2017/).

Supporters of online safety regulation often characterise the online world as a 'Wild West'[87] where the laws of the offline world do not apply and operators are careless of risks to their users. In reality, as a Law Commission report found in 2018, criminal laws apply equally, possibly more stringently, in the online world as offline. Furthermore, research conducted for the impact assessment found that, in general, the measures platforms had in place were already proportionate to the risk of potential online harm (i.e., higher risk platforms had many more protections in place than low risk platforms). Human and automated moderation was present across all risk categories of platforms. Most platforms already conduct risk assessments, set terms of service, have acceptable use policies, conduct both human and automated moderation, allow users to report harm, and have systems to handle complaints. This is far removed from a Wild West, and even if measures currently fall short of what politicians and campaigners would like to see, the case for heavy handed regulation, prescribing a legal model, and centralised concepts of 'harm' and safety to override all of the measures platforms already have in place is not well made.

A second evidential problem arises in connection with the question of whether the Bill will succeed in reducing harm. One point is that, as the Regulatory Policy Committee noted, the impact assessment does not address the likelihood of content moving to under-resourced or non-compliant platforms, or increased usage of VPNs to access overseas sites. In the particular case of harm caused by scientific misinformation, the Royal Society has produced important research. It highlighted that 'censoring or removing inaccurate, misleading and false content, whether it's shared unwittingly or deliberately, is not a silver bullet and may undermine the scientific process and public trust'.[88] That suggests that encouraging systematic removal of such information from Category 1 platforms could backfire by increasing mistrust and encouraging conspiracy theories to fester outside of large platforms.

The evaluation metrics set out in the impact assessment are supposed to set out how the success or otherwise of the legislation will be measured in post-implementation review. In this case, perhaps necessarily given the amount of work to be done by Ofcom in operationalising the legislation, the metrics are vague. Research on causal link monitoring that would allow ministers, MPs, the industry and others to determine what effect the legislation is having, and distinguish it from other causes or organic developments in the industry is still under way and due to be published in 2022.[89] The fact that the Bill is proceeding without an account of cause and effect, and without metrics that could be used to evaluate success or failure is concerning, as it is suggestive of rushed policymaking, based on preconceived ideas. The government's lack of attention to the question of evidence is suggestive that there is no strong evidence available. Given all of the potential costs to freedom of expression, privacy and innovation outlined above, the measures are disproportionate.

The failure of the Online Safety Bill's impact assessment to present satisfactory assessment criteria is not unique. In recognition of this general issue, the government's regulatory reform agenda includes improving post-implementation reviews and a commitment to 'thoroughly analyse

---

[87] For example, then Secretary of State for Digital, Culture, Media and Sport, Matt Hancock, declared in 2018 that 'the Wild West for tech companies is over' ('"Wild West" era for technology firms like Facebook and Google is over, minister declares', *The Telegraph*, 18 March 2018, https://www.telegraph.co.uk/technology/2018/03/18/wild-west-era-technology-firms-like-facebook-google-minister/). Damian Collins MP, a member of the Joint Committee on the draft Bill said in December 2021 'The Committee were unanimous in their conclusion that we need to call time on the Wild West online. What's illegal offline should be regulated online' ('Online Safety Bill: MPs and Peers call for new offences to tackle 'Wild West' online', Sky News, 14 December 2021, https://news.sky.com/story/online-safety-bill-mps-and-peers-call-for-new-offences-to-tackle-wild-west-online-12494892). For critique see 'Regulating online harms', House of Commons Library, 15 March 2022 (https://researchbriefings.files.parliament.uk/documents/CBP-8743/CBP-8743.pdf).

[88] 'The online information environment', The Royal Society, 19 January 2022 (https://royalsociety.org/topics-policy/projects/online-information-environment/).

[89] See page 107 of the impact assessment (https://publications.parliament.uk/pa/bills/cbill/58-02/0285/onlineimpact.pdf).

our interventions based on the outcomes they produce in the real world and where regulation does not achieve its objectives or does so at unacceptable cost, … ensure it is revised or removed'.[90] The impact assessment does not, in its current form, lay solid foundations for the desired improvements in evaluation of this legislation.

Conclusion

The Bill's scope, complexity and reach are breath-taking. It is 255 pages, an increase of 110 pages since the May 2021 draft, and 190 sections. This briefing has considered only certain aspects of the Bill but highlights how the duties it is proposed to impose on platforms threaten free speech and privacy. This briefing emphasises how the Bill threatens free speech and privacy while imposing immense regulatory burdens on platforms that will stifle innovation. These burdens seem very likely to lead to substantial overcompliance or, alternatively, to platforms leaving the UK. In the former case, free speech and privacy will be further compromised; in the latter, consumer choice and welfare damaged. And all this is undertaken in the absence of any convincing evidence of the need for regulation or of there being likely, measurable benefits.

[90] 'The Benefits of Brexit: How the UK is taking advantage of leaving the EU', HM Government, January 2022 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1054643/benefits-of-brexit.pdf).