SUBMISSION TO PARLIAMENT

Addressing automated, arbitrary, algorithmic censorship in the Online Safety Bill

June 2022

Contents

About Open Rights Group

Open Rights Group (ORG) is the leading UK-based digital campaigning organisation. We work to protect fundamental rights to privacy and free speech online, including data protection, the impacts of the use of data on vulnerable groups, and online surveillance. With over 20,000 active supporters, we are a grassroots organisation with local groups across the UK. We have worked on this Bill throughout the 'online harms' processes and consultations, and both Digital Economy Acts (2010 and 2017), accurately highlighting which parts of both DEAs would prove extraordinarily difficult to implement practically or fairly.

Addressing automated, arbitrary, algorithmic censorship

Our concerns and recommendations

We are concerned about the measures in this Bill to restrict online content using automated systems rooted in artificial intelligence and algorithmic processing. Our concern relates to arbitrary restrictions on lawful speech imposed by private companies. We outline our position with respect to automated content moderation on user-to-user services. We are concerned that some measures are a form of prior restraint. We have further concerns around the requirement for content removal on encrypted messaging services, with potentially dystopian outcomes (chat controls).

This is not merely about a tweet being taken down or a social media account suspended. It's about the way that the measures prescribed in this Bill, intended to remove one set of harms, will themselves create another form of harms. In doing so, they go to the very heart of our democracy and core British values of freedom of speech. It is Parliament's role to balance the competing interests.

We set out some robust safeguards for users that the Bill could incorporate. We would like to see clear and precise definitions of the content to be restricted, on the face of the Bill. We recommend *ex ante* and *ex post* procedural safeguards. Users should notified with a factual justification, including evidence, explaining why their content is restricted. They should have access to an effective appeals process, with the possibility for judicial redress and an effective remedy.

Summary

The Bill reflects a radical departure from the framework of law that currently governs Internet services. The effect of the Bill's provision is to create a mandate for the wide-scale monitoring of every social media post, and potentially every chat message too. This monitoring could only be conducted by means of artificial intelligence systems. In doing so, it automates the process of determining whether or not speech is lawful, and therefore, whether it should be censored.

This kind of 'general monitoring' has been forbidden in law to date. The imprecise and over-broad language in the Bill will be difficult for automated systems to interpret, and therefore arbitrary restrictions on users' lawful content are likely to increase, rather than go away, as the government claims. It is an interference with free speech rights that is incompatible with the right to freedom of expression under the Human Rights Act.

This Bill vastly over-reaches its remit. It imposes a statutory requirement to detect and remove illegal content, such as terrorism content and child sexual abuse material, as well as 23 other criminal offences listed on the face of the Bill, including assisting illegal immigration, firearms, financial services, and harassment. The mandate applies to social media platforms, search engines, and messaging systems. Moreover, it extends its scope outwards to thousands of small services that will have to pay a licence fee to Ofcom. It is literally asking private companies to enforce the law online on behalf of the State. Much of this scope has been added late and insufficiently scrutinised, including messaging systems, age verification measures and new "user choice" measures that impact anonymity.

The Bill requires the largest social media platforms to take restrictive measures against content that is not illegal but is either abusive, hateful or in some other way could cause harm (so-called 'legal but harmful'). The only way they can comply is with automated content moderation. The over-broad scope is likely to result in many false positives.

There is an implied requirement to monitor content on encrypted messaging services. This raises concerns about privacy rights. The policy aim is to tackle child sexual abuse material, but the effect will be a universal monitoring of chat messages, including to-do notes, photos and chains of contact. It is a form of bulk intercept, without suspicion or warrant. It is a spy tool in people's pockets that opens a Pandora's box for authoritarian rulers elsewhere who may seek to copy it for less benign purposes.

The requirement for age assurance was included at the last minute in the Bill now before Parliament. This is a new area of technology employing artificial intelligence (AI) , algorithmic processing and biometric profiling. AI still an emerging area and it is presumptive for Parliament to put this requirement on a statutory footing at this stage.

There is the concern around the powers granted to the Secretaries of State for DCMS and the Home Office who co-sponsor the Bill. They will not only define harmful speech to be restricted, but will have the power to make changes to the services in scope and the functions the regulated Internet services are required to comply with, as well as to strategically direct Ofcom and review its Codes of Practice.

We note that the government has recently signed the International Declaration on the Future of the Internet[1]. The Declaration issues a reminder of the importance of the Internet as an interconnected global communications system that is used by people use in their daily lives, and commits to a vision of a safe and trustworthy Internet that ensures the protection of human rights online. Parliament has a duty to balance the competing interests that are at stake in this Bill.

Automated censorship and arbitrary restrictions

This Bill relies on algorithmic processing in order achieve its aims, not only for content moderation but also in the mandate for automated age assessments (age verification). It is a blunt tool. An algorithm can only do what it is programmed to do. Precision is required. It is unclear how an algorithm could identify loosely-defined content as it currently stands in the Bill. The algorithm will over or under perform, according to how the requirement has been interpreted.

The Bill incentivises the taking down of content under threat of serious fines, and therefore it is likely that Internet services will be over-enthusiastic in applying restrictions, rather than give the benefit of the doubt. Moreover, the Bill requires providers to take measures 'if it is proportionate to do so [see for example Clause 9(4), 11(4) and 24(4)]. This suggests that Internet services (user-to-user and search) should determine the proportionality of their own measures as well as illegality of the content. This is somewhat concerning.

1 A Declaration for the Future of the Internet https://ec.europa.eu/commission/presscorner/detail/en/IP_22_2695

When we talk about arbitrary restrictions, we mean that content has been removed, or access denied, without explanation, or justification. This means that no evidence has been supplied to the user to justify the restriction, no statement given as to why is was necessary, and there is no clear link between the content restricted and the terms of service or the law. This is a frequent experience of users who have been subjected to restrictions by social media platforms. These removals will affect users who are acting lawfully, and in these cases, they would be a restriction on the freedom of expression of those users, as we outline below.

Content moderation systems

Automated content moderation systems will needed in order to comply with the Bill. This will apply to user-to-user services [Clause 9 (7) (Illegal content); Clause 11 (7) (content harmful to children); Clause 13 (content harmful to adults) and search services [Clause 24(4)]. Internet services may be required by Ofcom in a Code of Practice to install content moderation systems [Schedule 4(12)] or they may be compelled to do it by a Technology Notice issued by Ofcom (Clause 184(11)].

The Bill defines content moderation systems [Clause 184] in very limited way that is ill-suited to the legislative task: ' *technology, such as algorithms, keyword matching, image matching or image classification*'. It only applies to content regulated by the Bill, and excludes situations where content has been reported. This does not make sense and raises questions of due diligence with regard to the technology that the legislation is mandating.

Content moderation is the process by which online platforms determine whether or not text, images and videos may not be permitted on their systems and what action should be taken. It is more complex than the binary decision about whether to allow or take-down.[2] It involves defining and identifying the precise content, detecting it on the system, evaluating the actual files that have been located, and then determining the restrictive action from a menu of possible actions. The process of defining the content may refer to the platform's own rules, and to the law.

In the process of defining and evaluating, the platform may need to consider the context, which may include the type of account, the timing of the posting, and the words accompanying an image or a video.

Content moderation systems use artificial intelligence and are driven by algorithms. These systems are trained to recognise fingerprints – hashes - of images. They scan

2 CDT Outside Looking In: Approaches to Content Moderation in End-to-end Encrypted Systems August 2021

massive databases of these hashes in order to seek a match against the images on the platform (a technique known a predictive hashing). Some of these databases have been built as a shared industry initiative. For example, the database of terrorism and violent extremism content operated by the Global Internet Forum to Counter Terrorism, which is funded by the big four global online platforms. Others are developed by the large online platforms independently such as Facebook's database of non-consensually shared sexual images. [3] Alternatively, content moderation systems may analyse metadata or user behaviour (which includes frequency of messages or posts, reports from other users) combined with machine learning to recognise characteristics of content.

Content moderation systems determine the restrictive action to be taken. This may be an apparently straightforward take-down of an individual piece of content, but it would false to think that was the only option. They can suspend a user's account or restrict it in some other way, and they can feed into recommender systems in order to demote or suppress certain content (shadow ban). These restrictive actions are listed in [Clause 13(4) c] as possible treatments for 'content harmful to adults'. It's worth noting that content moderation decisions could fall under the remit of Article 22 of the General Data Protection Regulation (GDPR) that gives users the right to know about an automated decision that affects them. [4]

On search services the restrictive actions will entail demoting content in search listings. Search results show content from all over the web. will therefore affect all websites, even those that are not in scope. This raises concerns that websites could become hidden with little or no redress, as they possibly would not even know what had happened.

Arbitrary restrictions

The scale of deployment of content moderation systems that will required to comply with Bill, raises serious risks for freedom of expression, and the possibility for arbitrary restrictions. There are huge question marks around how content moderation systems could define and identify the content that the Bill seeks to address. Automated systems need precise definitions of the content to be restricted, which the Bill does not provide. Throughout the Bill, the definitions of the different types of

3 Ibid CDT

4 Future of Privacy Forum: Automated Decision-making under the GDPR, Case 24 https://fpf.org/blog/fpf-report-automated-decision-making-under-the-gdpr-a-comprehensive-case-law-analysis/

regulated content are vague and wide open to interpretation. The systems cannot make determinations based on context, and they cannot determine intention. Therefore it's likely that Internet services would over-play the restrictions to avoid getting a fine.

For example, how should these systems identify 'harmful' or what is meant by terms such as 'journalistic content' and 'content of democratic importance'. When it comes to illegal content, there will be similar difficulties. The offences defined on the face of the Bill describe criminal actions committed by people. The determination of whether content is illegal, requires context, which may include the intent of an individual. The vague and imprecise language in this Bill means that users cannot foresee whether or not the content they post will comply with the law.

It is likely to lead to uncertainty, both for users, and for Internet services (user to ser and search) who will not know specifically which content complies or does not comply for the purposes of the restrictive measures. Individual services will put a different nuance on the interpretation and will programme their algorithms accordingly. The result will be more content than necessary being censored.

The Bill gives Internet services enormous latitude to take restrictive actions on the basis that they '*reasonably consider*' content to be illegal or harmful. Overall, there is no requirement to provide a factual justification or supply evidence, The notion of 'harmful' content is itself nebulous, and will be dependent on the descriptions provided by the Secretary of State in secondary legislation that will be laid before Parliament after the Bill is passed. It would establish a new notion of speech that can be censored for which there is no offline equivalent.

The underlying concept of *'harm*' is defined as '*psychological harm amounting to at least serious distress'*. This begs a follow up question as to what is *'serious distress*'. Moreover, there is a catch-all where Internet services are expected to restrict content where they '*reasonably consider*' there is a '*material risk of significant harm to an appreciable number of adults in the United Kingdom'.* This could be just about anything identify as being harmful, and could be based on the personal views of the Secretary of State.

An often overlooked notion in this Bill is that of Clause 11(5) '*non designated content that is harmful to children*' (or simply '*other content harmful to children*' in Clause 11(3)(b)). This seemingly gives Internet services *carte blanche* to apply restrictions with no statutory guidelines at all. We can expect many arbitrary decisions.

Internet services (Category 1, user-to-user) may choose how they restrict content harmful to adults, as long as they specify the methods in their terms and conditions. The Bill gives them options to either take down content, or limit its recommendation or promotion in feeds and timelines (shadow ban), or restrict the users' access to the content.

If they take the measures as prescribed in a Code of Practice that will be drawn up by Ofcom, they will be deemed to have complied with the Bill. From their viewpoint, this may seem like a safer position. However, for users, it represents a precarious outcome where they have continual uncertainty and are at the ever-present mercy of the content moderation algorithms.

One might be forgiven for thinking there was less of an issue with illegal content because it is 'defined' on the face of the Bill. The Bill identifies terrorism content and child sexual abuse material, as priority illegal content. It cites the relevant criminal offences that would apply. In Schedule 7, the Bill lists 23 separate criminal offences, including assisting illegal immigration, sexual exploitation and harassment, as 'priority illegal content'.

However, it does not provide any specific definitions of how the content should be identified. This is problematic. Illegal content is to be interpreted as '*use of the words, images, speech or sounds*' that constitute an offence [Clause 41(3)]. The Joint Committee on Human Rights has already identified the difficulties, asking how a provider of user-to-user services would identify an offence under Section 5 of the Public Order Act 1986 in a social media post? [5] Similar questions apply to every one of the 23 offences in Schedule 7.

The Independent Reviewer of Terrorism Legislation[6] suggests that the definition of 'terrorism content' in Clause 52(2) and 52(5) is inadequate for a determination to be made that the content itself is illegal because it leaves out the mental element or intention, as well as the possibility that a defence is available. He concludes that the uncertainty created by the weak definition creates an uncertainty around what might or might not be 'terrorist content' and that uncertainty is likely to result in either too little moderation or over-zealous removals. We note that the United Nations Special Rapporteurs said in a recent report that *'it is difficult to determine with reasonable certainty what kind of conduct online would be considered terrorism',* and that the use

5 Joint Committee on Human Rights, letter from Rt Hon Harriet Harman MP to Secretary of State for DCMS, 19 May 2022

6 Independent Reviewer of Terrorism Legislation, Missing Pieces: A Note on Terrorism Legislation in the Online Safety Bill

of artificial intelligence technologies to prevent its dissemination may curb free speech.[7]

Why prevent has a special meaning (prior restraint)

The requirement for online platforms to 'prevent' certain regulated content from appearing online at all, is a form of prior restraint. It breaks from the age-old principle in English law that forbids censorship before publication.

The word 'prevent' occurs in [Clause 9(3)(a) *prevent individuals from encountering priority illegal content by means of the service*]. It requires Internet services to scan content as it is being uploaded by users (the so-called upload filter), and then to seek out and make a judgement as to the illegality that content. Where it finds priority illegal content, it should remove it and thereby 'prevent' it appearing on the platform. Priority illegal content is defined as terrorism content, child sexual abuse material and 23 additional criminal offences [Clause 52 (7) and Schedule 7].

An upload filter constitutes a form of prior restraint[8]. It represents a particularly severe restriction on the right to freedom of expression and is incompatible with the European Convention on Human Rights[9] In this instance, it is deeply problematic because the Internet services are being asked to restrict content that they *'reasonably consider'* to be illegal, with no requirement to examine evidence. The Bill does not state how the offences would be interpreted in terms of content on social media posts. The decision would be taken by algorithms which cannot take into the account the context. We recommend that this measure is removed from the Bill.

Hands, face, age-gate

The word 'prevent' occurs again in Clause 11(3)(a), referring to preventing children from accessing harmful content. However, the meaning in terms of the technology to be applied appears to be different, because the Clause specifies '*age verification or another means of age assurance'*. Age verification is usually used to refer to systems that provide an accurate assessment of a person's age by verifying their passport or other form of ID. However, the term age assurance can also refer to systems that estimate a person's age and place them within an age bracket, in order to establish services tailored to the needs of specific age groups. They can provide a binary determination as to whether someone is or is not an adult.

---

7 Mandates of the Special Rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, United Nations, OL OTH 229/21 21 October 2021

8 CJEU upholds Article 17, but not in the form (most) Member States imagined - Kluwer Copyright Blog http://copyrightblog.kluweriplaw.com/2022/04/28/cjeu-upholds-article-17-but-not-in-the-form-most-member-states-imagined/

9 Comparative Study on Blocking, Filtering and Take-down of Illegal Internet Content, Council of Europe, 20 December 2015

These systems estimate age using algorithmic processing of personal data, including biometric data, which may include hand or facial images, handwriting, or voice. They may also use techniques such as behavioural analysis or profiling, using personal data gathered from activity online; or they may use vision-based analysis that estimates someone's age from an image. They use artificial intelligence techniques to analyse the data, and, according to a report by the Information Commissioner's Office (ICO)[10] there is little evidence for their effectiveness or accuracy.

All of these techniques are privacy intrusive and are high risk from a data protection perspective. They would most likely be operated by third-party service providers. Strict data protection safeguards, including data minimisation and purpose limitation would need to be imposed. Profiling must be proportionate to the risks to children. As the ICO report highlights, there are further risks from algorithmic bias, and some of the techniques they use are still new and may require further testing.

User-to-user and search services are required to do an of the risk that children will use their services. There are outstanding questions as to how these systems would operate in practice. Service providers may have a Hobson's choice to either age-gate their platforms, or sanitise the content to a level suitable for all age groups.

This mandate was only included in the Bill at the last minute. There is a risk of legal challenges from industry stakeholders, who have previously taken the government to court about its decision not to proceed with age verification for websites hosting pornography[11]. However, the governance of artificial intelligence is the subject of international discussion, and techniques such as biometric profiling are flagged as high risk, and we believe that more due diligence is needed on the operation of these technologies.

Ministerial powers to interfere with speech

The Bill gives unprecedented powers to the Bill's joint sponsors, DCMS and the Home Office to define speech. It is a power incompatible with freedom of expression, under Article 10. The Secretary of State will have powers to make regulations to define content harmful to adults [54 (2) and 54 (3)] as well as content harmful to children [53 (2) and 53 (3)] , as well as powers to amend the specification of illegal content in Schedules 5,6, and 7. [Clause 176].

---

10 Information Commissioner's Opinion : Age Assurance for the Children's Code 14 October 2021

11 The Guardian, UK government faces action over lack of age checks on adult sites https://www.theguardian.com/society/2021/may/05/uk-government-faces-action-over-lack-of-age-checks-on-pornography-websites

This is a dangerous move. It offers considerable leeway to a less benign government that may wish to introduce draconian and authoritarian-style censorship. Harm could mean whatever the Secretary of State chooses it to mean. To see how easily this can be done, one only has to look at how Russia has enforced a rule on the sharing of 'fake news' about the war in Ukraine, adopted in March this year, in order to suppress public debate.[12]

These powers work together with additional powers to direct Ofcom to comply with public policy and to modify its Codes of Practice [Clause 40 (1 (a) and 40 (7)] and to amend the scope of the Bill after it becomes law. All of these powers would exacerbate the arbitrary removals of lawful content and could be achieved through a bypass of Parliamentary scrutiny to impose changes through Secondary legislation (Henry VIII powers).

Chat controls and the spy in the your pocket

The requirement for private messaging services to scan and moderate content, raises deep concerns. The technology that would be required would be like putting a spy in everyone's pocket, and it opens a Pandora's box for authoritarian-style censorship.

Whilst the measure is intended to address a heinous crime, there must also be checks and balances to ensure that it does not create other harms across the whole population.

Private messaging services have been brought within scope of the Bill by sleight of hand in the drafting. The Bill establishes a category of "user- to-user service". This is a new legal category for the purposes of this legislation. It is service that enables content to be uploaded, shared or encountered by users, and as such it could equally describe a social media or messaging service. 'Content' is anything communicated 'publicly or privately' [Clause 189 interpretation – "Content"]. This follows through into the Clauses that provide for 'regulated content' and the requirements for Internet services to comply with.

The requirement for private messaging services to moderate content is not explicit, but is implied for illegal content, . It significantly expands the scope of the Bill because private messaging services have quite different characteristics from the public social media platforms, that are the primary target of the measures in this legislation. In particular, users on private messaging services mostly communicate one-to-one or in

12 Amnesty International, Russia: Authorities launch witch-hunt to catch anyone sharing anti-war views https://www.amnesty.org/en/latest/news/2022/03/russia-authorities-launch-witch-hunt-to-catch-anyone-sharing-anti-war-views/

small groups.

Content moderation on private messaging services would operate in using similar processes to those outlined above – it would seek out, identify and evaluate the content to be restricted, and then apply a restrictive action, according to a pre-programmed set of rules. However, there is a difficulty because the content is encrypted. Technically, it is known as end-to-end encryption (often written e2ee for short).

"End-to-end" means that the content is encrypted from the moment it leaves your smart phone, to the moment someone else reads it on their phone. All along the way, the message travels as data bits and bytes, over wires and radio frequencies, through various transmitters and routers, and possibly various other services, and no-one can read it. That's important. It's a guarantee that it is confidential, that it is from the person it says it from, and that it is the message that was intended to be sent – no-one could have tampered with it.

Content moderation systems can only scan unencrypted content. That means clear text or images that the system can read. In order to be able to do this, they need to either scan the content on the server, meaning that they have to break the encryption. This compromises the guarantees of integrity, authenticity and confidentiality, and introduces vulnerabilities into the system. these vulnerabilities could be a way for bad actors to hack into the system.

The other way is to scan the content before it becomes encrypted on the user's smartphone. This is known as client-side scanning. Technically, it is deemed not be breach the encryption, but in fact it has the potential to be far more dangerous[13]. The scanner now resides in the user's smartphone, and it is not just monitoring for the illegal content but it enables the remote searching of not only chats, but a wide range of content on phone, including to-do lists, personal notes, and chains of contacts. It creates a means of surveillance of people's intimate communications and thoughts.

It is understood that the UK government wants to use client-side scanning and could ask Internet services to implement it under the Technology Notice provisions in the Bill. These provisions pave the way for Ofcom 'verified' technology to be mandated. 'Verified' systems will operate to standards approved by the Secretary of State, likely to be the Home Office. Why this matters is that private messaging services are becoming the dominant form of communication between citizens. They are used around the

13 Abelson, H and Anderson, R, et al Bugs in our Pockets: The risks of client-side scanning https://arxiv.org/abs/2110.07450

clock in daily life at home, at work, at leisure. People use private messaging services to keep in touch with friends and family. These services have replaced the old-fashioned phone in business, and even in Westminster!

Client-side scanning would be a vastly disproportionate interference with the privacy of the majority law abiding population. Rather like having a spy in your pocket. It is questionable whether this Bill provides the correct legal basis for imposing this requirement on encrypted messaging services. It has the characteristics of bulk intercept which is addressed under the Investigatory Powers Act.

User Dis- Empowerment

We have some concerns about the 'user empowerment' provision in Clause 14. The policy goal of these measures is not clear, nor why it is needed to be on a statutory footing. The government's factsheet states '*Women will have more decision-making over who can communicate with them and what kind of content they see on major platforms. This will strengthen the protections against anonymous online abuse.'* However, the Clause applies to all adult users (not just female users) and it is a confusion of targeting non-verified users on the one hand, and alerts to harmful content on the other. It references an unspecified verification process in Clause 57, leaving wide open the option for Internet services to determine what 'verified' means. One unconfirmed interpretation is that the government wants all users of user-to-user services to be identified, and importantly, not anonymous. If this is the policy aim, then it should be clear on the face of the Bill so that Parliament can scrutinise it. We would respectfully suggest that verification of someone's identity is not a guarantee of their integrity, and bad actors or spreaders of disinformation may also hold verified accounts. How will this help vulnerable women?

Rights for online speech

A flaw in this Bill is that if does not recognise that users have positive rights to free speech, nor does it recognise that interference with free speech rights could occur. This matters because there is a long tradition of free speech in British democracy. In 2022, people exercise their free speech rights online and on social media platforms or user-to-user services.

Free speech rights have been hollowed out and reduced to mere contractual matters between the draft Bill and the Bill as introduced to the House of Commons. The Bill merely asks to '*have regard to the importance of protecting users rights to freedom of*

expression within the law' [Clause 19(2) and 29(2)].

Likewise, the right to privacy is no longer acknowledged as a 'right' . The Bill [Clause 19(3) merely refers to a '*statutory provision or rule of law concerning privacy'*. It is not clear what this means. However, we would draw attention to an important provision in the GDPR is the right of individuals to know how an automated decision that affects them was taken (Article 22 GDPR). Privacy rights also relate to surveillance activities. Blanket measures to monitor across an entire population, such as those proposed to monitor communications on private encrypted services, are not compliant with human rights law.

Whether acting as speakers or as recipients of information, people using Internet services have positive rights.

In 2022, this is the vast majority of the UK population. Free speech is a right that applies to everyone who uses online platforms and Internet services, under the Human Rights Act and the European Convention on Human Rights, Article10. It is a two-way right to speak and to receive information, without interference from public authorities. The State has a duty to guarantee these rights for *all* users against arbitrary interference or restrictions. However, it is not an absolute right and may be restricted under certain strict conditions, where the measures must be balanced against any interference with the rights of people who are not the target.

Automated content moderation systems engage Article 10 because their restrictive actions represent an interference with freedom of expression. Our concern is for those instances where the restrictive measures in the Bill represent an unjustified interference with lawful speech. Where speech rights are going to be restricted, the restrictions themselves must meet certain legal criteria. They must be strictly necessary to meet a legitimate aim, and they must be proportional to that aim. Any measures taken to implement restrictions on free speech must be the least restrictive needed to achieve that aim. They must be targeted and defined as narrowly as possible. The quality of the law is also important. It must be clear and precise and unambiguous so that people know what they are not supposed to do (or not do) and can adjust their behaviour accordingly.

All of this applies to speech online in the same way as to speech offline. The way that it applies has been determined in case law in the UK courts. For example, the need for the least intrusive measures to be chosen, and for the measures to be narrowly prescribed, has been determined in UK courts. The exact locations of the content to be restricted, such as URLs, should be provided and the restriction must be limited to that

content.

Users who are restricted – for example, their content has been taken down or their account terminated – should have the right to a fair hearing if they believe their content was lawful or did comply with the terms of the Internet service. This may be a judicial or administrative hearing. They are also entitled to an effective remedy. These rights should be statutory, on the face of the Bill. As it stands, this is a serious omission.

Procedural safeguards and effective remedies

Our recommendations for procedural safeguards are based international standards such as the Council of Europe Recommendation on Internet Freedom [CM/Rec(2016)5][14]. We have also based them on our own more immediate experience of dealing with users who have had restrictions placed on their social media accounts and content posts. We would like to take this opportunity to share some of those experiences.

User experiences of restrictive content moderation measures

We have seen examples of users being restricted without warning or explanation. Social media platforms use a variety of restrictive actions which are not limited to content removal or accounts suspension. Users generally don't understand why the restriction has been imposed, or how they can appeal. They often do not know what they should be appealing. They have either not received a notice, or if they have, they don't understand it because it tend to be written in jargon that's used internally, but means nothing to anyone outside the company.

The most recent report we received was in late May when a user's Twitter account was closed without notice or warning. The user had no idea why, shrugged it off and started a new account. This is a typical reaction. There is an inconsistency in notifications. Sometimes users receive one, and sometimes not. However, it is a common factor that the restrictions mostly appear to be arbitrary. Sometimes they relate to the takedown of a post that has been shared and the user was likely to have been one of many targeted. Even though the user did not upload the content, and only 'shared' (forwarded) it, they are sanctioned, and the user will get a 'strike' for that takedown. 'Strikes' are then added up and used to determine whether or not to impose other restrictions, such as suspending the account.

The way in which users' posts can be mis-interpreted by automated systems and why context matters was illustrated in a case ruling from the Facebook Oversight Board. It

14 Council of Europe Recommendation on Internet Freedom [CM/Rec(2016)5]

concerned a user who had made a post involving a quote attributed to Josef Goebbels[15]. The post was restricted under Facebook's Dangerous Individuals and Organisations policy which is used for terrorism content. The Oversight Board ruling is instructive: the context of the post, notably the comments below it, were sufficient to show that the user was making a political comment and did not support terrorism. The Board found that the user was not told which Community Standard he had violated, and there was a gap between Facebook's public and non-public rules. The Board required reinstatement of the post. In a separate case, a user's post of a Guardian article discussing the case of Shamima Begum, was restricted. It was shared on Facebook by the journalist Jon Danzig. The post was taken down, and his account was banned from posting for 30 days, also under the 'Dangerous Individuals' policy.[16]

We have seen several examples of what's known as a shadow ban – this is a specific term for content that is hidden or de-prioritised without informing the user. Shadow bans are the intended meaning of Clause 13 (4)(c) *limiting the recommendation or promotion of the content*. This wording appears to be mis-understood by a number of stakeholders. Shadow bans are created by suppressing the distribution or promotion of posts in users' timelines or feeds. Those feeds are operated algorithmically by recommender systems. It's possible to suppress individual posts or types of posts, or an entire account this way. The user is unaware of the suppression until they notice the number of clicks falling, and it can reduce the readership numbers to the extent where the effect of it is not much different from a total ban or removal.

Procedural safeguards against arbitrary take-downs

Currently the Bill fails to recognise at all that there could be any interference with free speech rights. Safeguards against interference are absent, beyond giving lip service to a complaints process. Based on our experience, we believe that robust *ex ante* and *ex post* procedural safeguards are needed.

We believe that the kinds of arbitrary restrictions that we have seen will increase with this Bill. As we have stated above, we think this will occur as a direct consequence of the vague and indeterminate language, the imprecise definitions of the content to be restricted and the over-broad discretion give to the Internet services to interpret these definitions when they code their content moderation systems. An obvious move would be to introduce an effective appeals process. This could be incorporated by amending the Bill's complaints procedures in Clause 18 (user-to-user services and 24

15 Facebook Oversight Board Back to decisions Case decision 2020-005-FB-UA https://www.oversightboard.com/decision/FB-2RDRCAVQ

16 Jon Danzig, Facebook backs down after wrongly banning me, 1 March 2021 https://www.linkedin.com/pulse/facebook-backs-down-after-wrongly-banning-me-jon-danzig/

(search services) to create a statutory right of appeal and effective remedy.

The Bill as it stands requires only that a complaints procedure is established but does not consider how it would operate. Interestingly, the Bill appears to allow for some complaints, including about shadow bans, regardless of whether they address regulated content. Clause 18 (4) (e) states that users may complain if : *the use of proactive technology on the service results in that content being taken down, given a lower priority in other users' feeds or being otherwise restricted, and the user considers that the proactive technology has been used in a way not contemplated by, or in breach of, the terms of service.*

As mentioned above, users frequently don't appeal because they don't trust the process. In order to balance the measures in this Bill, it is important that users can have trust in the way that decisions about their speech are being made. The operation of the appeals process should be on a statutory footing. It should be possible for appeals electronically and free of charge. If the appeal found in favour of the user, the restrictions should be swiftly removed. The Internet service should be required to perform this process diligently, and to give an explanation of the outcome.

Before the restriction is imposed, or at least at the time of imposition, users should be notified about the decision. They should be told the specific content that is to be removed, along with a clear and specific statement of reasons for that decision, including the rule or the law that was used[17]. The notification should include information on how the decision was taken (and if taken by automated systems). It should provide the grounds for illegality with evidence, or why it is harmful, or why it does not comply with the platform terms and conditions. The notification should say how user may appeal the decision, and the deadline to lodge the appeal.

Ideally, the assessment of illegality should be made by a court or public authority, and not by the Internet service. Users should also have a statutory right to judicial redress, in compliance with human rights law, and be informed of this right. If this process were put onto a statutory footing, there would be no need for the separate protections for journalistic content or content of democratic importance [Clauses 15 and 16], since *all* users would have the right to use these processes.

17 See Footnote 15 Facebook Oversight Board