ONLINE SAFETY BILL – SUBMISSION TO PUBLIC BILL COMMITTEE

26 May 2022

SUBMISSION BY GRAHAM SMITH

Graham Smith is a solicitor in private practice in London. He is the editor and main author of the legal textbook *Internet Law and Regulation* (Sweet & Maxwell, 5th edition Dec 2019). He also writes the Cyberleagle blog.

This submission is made in Mr Smith's personal capacity. The views expressed are not attributable to the law firm at which he works or to any of its clients.

Contents

1. INTRODUCTION

1.1 This submission focuses on the illegality duties in **Clause 9** of the Bill, tied to the definition of illegal content in **Clause 52**. Until recently these duties have received comparatively little attention, perhaps due to a perception that they raise no particularly difficult issues and will attract universal consensus:

> "I think there is absolute unanimity that the Bill's position on that is the right position: if it is illegal offline it is illegal online and there should be a duty on social media firms to stop it happening. There is agreement on that." (Chris Philp MP, Minister for Technology and the Digital Economy, Evidence to the Commons DCMS Sub-Committee on Online Harms and Disinformation, 1 Feb 2022.)

1.2 If that was the perception, it ought by now to have changed. The Independent Reviewer of Terrorism Legislation has recently pointed out problems with the interaction of the illegality duty with terrorism offences[1]. (Similar problems arise with other in-scope offences). The Joint Parliamentary Human Rights Committee has written to the Secretary of State raising questions about the illegality duty in relation to Article 10 of the European Convention on Human Rights (ECHR)[2]. These issues deserve serious consideration.

1.3 The illegality duties would apply not only to 'social media firms' (the Minister's phrase), but to all in-scope U2U service providers[3]. According to the Impact Assessment those c. 25,000 businesses and organisations (including over 500 civil society organisations) would encompass 20,200 micro-businesses, 1,200 small businesses, 2,900 medium businesses and 700 large businesses[4]. They would include, for example, an MP who included a discussion forum on their constituency app or website.

2. THE ISSUES

2.1 The substantive illegality duties in Clause 9(3) require a platform to use proportionate systems and processes designed to prevent individuals from encountering priority illegal content; minimise the length of time for which any priority illegal content is present; and where the provider becomes aware of the presence of any illegal content,

---

[1] *Missing Pieces: A Note on Terrorism Legislation in the Online Safety Bill*, Independent Reviewer of Terrorism Legislation (Jonathan Hall QC), 20 April 2022 (https://terrorismlegislationreviewer.independent.gov.uk/missing-pieces-terrorism-legislation-and-the-online-safety-bill/)

[2] Letter, 19 May 2022. (https://committees.parliament.uk/publications/22323/documents/165077/default/)

[3] Less extensive illegality duties would apply to search engines. Many of the issues identified in this submission would also be relevant to those duties, since like the U2U duties they are founded on the definition of illegal content in Clause 52(3).

[4] Impact Assessment, 31 January 2022, para 109 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1061265/Online_Safety_Bill_impact_assessment.pdf).

swiftly to take it down. It is implicit in these duties that the platform will seek to determine what user content is and is not illegal.

2.2 The Terrorism Legislation Reviewer and the Human Rights Committee both raise, in different ways, an overarching question: **how does the Bill require a platform to perform the task of adjudging the legality of user content**?

2.3 That question resolves into three parts:

2.3.1 **Interpretation of the Bill**. According to what criteria does the Bill require a platform to adjudge legality?

2.3.2 **What information** does the Bill contemplate as being available to a platform when making that judgment?

2.3.3 **What technical or human processes** does the Bill contemplate that a platform is capable of bringing to bear on that information?

2.4 The third part, at least for the proactive obligations imposed by Clause 9(3)(a) and (b), brings automated monitoring and filtering techniques into consideration[5]. The Human Rights Committee raised specific questions about how automated systems would determine on which side of the line (legal or illegal) content falls.

2.5 The answers to these questions bear heavily on the **compatibility of the Bill with the fundamental right of freedom of expression**, most significantly as protected by Article 10 ECHR but also in some respects (such as insistence that the law be clear and certain) as a long standing domestic constitutional principle.

3. ADJUDGING ILLEGALITY – INTERPRETATION OF THE BILL

3.1 **According to what criteria does the Bill require a platform to adjudge illegality?** This raises two main issues: which ingredients of illegality does a platform have to adjudge are present; and to what level of confidence?

3.2 **Ingredients of illegality** The Terrorism Legislation Reviewer has pointed out the conceptual problem with describing an item of content as being of itself either legal or illegal:

> "Under terrorism legislation content itself can never 'amount to' an offence. The commission of offences requires conduct by a person or people."

Although directed specifically to terrorism offences, his observation applies to criminal offences generally.

3.3 The Bill, at Clause 52(3), tries to solve this conundrum by providing that content amounts to an offence if specified conduct in relation to that content - use, possession, viewing, accessing, publication or dissemination - would amount to an offence. The

[5] Proactive detection and filtering can be performed manually at small volumes. However, the Bill and the Impact Assessment contemplate significant use of technology for this purpose. Further, many of the issues discussed in this paper in the context of automated systems (such as capability of determining illegality merely by examining the contents of a post) apply also to human moderation.

intention appears to be to span the gamut of different ways in which in-scope criminal offences are framed.

3.4 The problem highlighted by the Terrorism Legislation Reviewer is that the Bill does not go on to make clear whether, when determining whether an item of content amounts to a given offence, other ingredients of that offence in addition to the conduct specified in Clause 52(3) are to be taken into account (and if so how).

3.5 In the case of terrorism offences that is a particular concern, since two ingredients other than conduct - the intention of the defendant and the availability of a defence (such as reasonable excuse) – do the heavy lifting in keeping the scope of the offence within appropriate bounds. The Terrorism Legislation Reviewer commented:

> "Intention, and the absence of any defence, lie at the heart of terrorism offending".

3.6 He observed that:

> "it cannot be the case that where content is published etc. which might result in a terrorist offence being committed, it should be *assumed* that the mental element is present, and that no defence is available. Otherwise much lawful content online would "amount to" a terrorist offence [for the purposes of the illegality duty]."

3.7 He then gave five examples of unexceptional online behaviour, ranging from uploading a photo of Buckingham Palace to soliciting funds on the internet, which on the basis of such assumptions would be caught by the illegality duty. He concluded that:

> "it would be unsupportable to require assumptions about mental element and lack of defence. But this leaves the question of intention and defence unaddressed."

3.8 A similar issue arises with any in-scope offence that contains a mental element and provides defences. Its significance is greatest where the conduct element is drawn in broad terms and the heavy lifting in terms of appropriately limiting the scope of the offence is done by the mental element and available defences[6]. Three offences provide some illustrations:

3.8.1 The new **harmful communications offence** (clause 150 of the Bill). The Law Commission stressed that in its view the compatibility of the offence with freedom of expression rested on the mental element (intention to cause harm to a likely audience) and the requirement for the prosecution to prove,

[6] It should also be emphasised that the question for a platform seeking to discharge its duty under the Bill is not 'Would the CPS prosecute?', but 'Is this content within the letter of the statute?' Mitigation of broadly drawn offences through the exercise of prosecutorial discretion is not a relevant consideration. Cf para 61 of the Opinion of Gavin Millar QC obtained by Index on Censorship in May 2022 (https://www.indexoncensorship.org/2022/05/online-safety-bill-will-significantly-curtail-freedom-of-expression/), highlighting the difficulty that a prosecutor would face in making correct decisions under S.13(1A) of the Terrorism Act 2000, suggesting that the chances of a platform making the 'right' decision in such cases is remote, and that the other priority offences are similarly difficult to judge.

beyond reasonable doubt, that there was no reasonable excuse for sending the message[7]. However, it is unclear how burden and standard of proof in the context of a criminal prosecution might translate (if at all) into the Bill's illegality duty. Burden of proof aside, the substance of the 'no reasonable excuse' ingredient is identical to that of a defence of reasonable excuse.

3.8.2 The newly codified **public nuisance offence** in S.78 of the Police, Crime, Sentencing and Courts Act 2022. The relevant conduct element of the offence is 'an act' that creates a risk of, or causes, serious harm to the public or a section of the public. 'Serious harm' includes serious distress, serious annoyance or serious inconvenience[8]. The requisite mental element is intention that the act would have that consequence or recklessness as to whether it would have such consequence. It is a defence to prove reasonable excuse for the act.

The Law Commission has discussed how the statutory public nuisance offence's common law predecessor could have applied to online communications[9]. There is no reason to believe that the new offence, which is couched in technology-neutral terms, would not also do so[10].

3.8.3 The **harassment offence** under the Protection from Harassment Act 1997, which is designated under the Bill as a priority offence. It contains a broad course of conduct element (in which 'harassment' is left undefined), a mental element that the defendant 'knows or ought to know [that the course of

[7] "However, by requiring proof that the sender intended to cause harm, genuine political commentary will be excluded from scope by virtue not only of the reasonable excuse element but also by virtue of the fault element." ([2.10]); "we have built Article 10 protection into the offence, in requiring proof that the defendant lacked reasonable excuse (and it is this test which imports consideration of political debate)" ([2.180]) The Law Commission set great store by lack of reasonable excuse being framed as an ingredient of the offence rather than as a defence of reasonable excuse: "our proposal was that the prosecution should be required to prove as part of the offence that the defendant lacked a reasonable excuse – and that needed to be proven beyond reasonable doubt. It was for this reason primarily that we considered the offence complied well with Article 10 ECHR. Unless the court was sure that the defendant lacked a reasonable excuse, considering also whether it was the communication was or was intended as a contribution to the public interest, the defendant would not be found guilty" ([2.204]-[2.205]).

[8] The offence would therefore fall within the definition of offences for which an individual is an intended victim under Clause 52(4)(d). At present the offence has not been designated as a priority offence.

[9] *Scoping Report on Abusive and Offensive Online Communications*, November 2018 ([8.136] – [8.143]).

[10] "The new statutory offence of public nuisance will cover the same conduct as the existing common law offence of public nuisance." Home Office: *Police, Crime, Sentencing and Courts Bill 2022: protest powers factsheet*, [4.9] https://www.gov.uk/government/publications/police-crime-sentencing-and-courts-bill-2021-factsheets/police-crime-sentencing-and-courts-bill-2021-protest-powers-factsheet

conduct] amounts to harassment of the other'[11] and a reasonableness defence[12].

3.9 **Level of confidence** To what standard does a platform tasked with adjudging legality have to be satisfied that an item of content is illegal? Does, for instance, it have to be satisfied beyond reasonable doubt (as would a criminal court) that the content amounts to a criminal offence? Or does some lower standard apply, and if so what?[13]

3.10 In the draft Bill the platform's illegality duty was triggered if it had reasonable grounds to believe that the content was illegal. That provision, which formed part of the definition of illegal content, is no longer present. As the Terrorism Legislation Reviewer has observed[14], content now either amounts to an offence or it does not. The Bill does not now expressly stipulate any standard to which the platform has to be satisfied of illegality[15].

3.11 **The interpretation issue** stems from the contradiction between S.52(3), which on the face of it appears to assume that the mere act of publishing etc an item of content will constitute an offence, and in-scope offences themselves (including priority offences listed in Schedule 7) for which more than that is required (whether intention, additional ingredients, or both) and for which defences exist.

3.12 Significant questions of interpretation arise both as to the elements of an offence that a platform has to consider when adjudging illegality and as to the standard to which it has to be satisfied of the existence of those elements. These are important in their own right, as a matter of clarity of the Bill. They also have consequences that flow through to the second and third limbs of the overarching question of how a platform should go about adjudging the legality of user content.

4. ADJUDGING ILLEGALITY – CONTEMPLATED AVAILABILITY OF INFORMATION

4.1 **What information** does the Bill contemplate will be available to a platform when adjudging legality?

4.2 That in part depends on the answers to the interpretation questions already discussed. If, for instance, intention and available defences (such as reasonable excuse) are meant

---

[11] Whilst the mental element for the harassment offence is framed in more objective terms than the harmful communications and the public nuisance offence, how it may relate to Clause 52(3) of the Bill is no clearer than for those offences.

[12] The significant problems with applying the Bill's illegality duty to the harassment offence are discussed in detail here: Cyberleagle, *The draft Online Safety Bill concretised* https://www.cyberleagle.com/2021/11/the-draft-online-safety-bill-concretised.html.

[13] The lower the standard, the more legal content the Bill would require a platform to sweep up and suppress. That has implications for ECHR compliance, discussed further in Section 6.

[14] *Missing Pieces*, at [10].

[15] The Opinion obtained by Index on Censorship from Gavin Millar QC argues that the standard remains one of reasonable grounds to believe, as a consequence of Clauses 9(5) and 9(6). Those impose duties to apply terms and conditions consistently in relation to content that the platform 'reasonably considers' to be illegal. However, the substantive removal duties under Clause 9(3), which contain no similar provision, are self-standing and sit alongside the terms and conditions duties. It is at best unclear how these provisions are meant to interact with each other.

to form part of the assessment of legality, the Bill must contemplate that information about such matters is available to the platform. Even if intention and available defences are meant to be left out of account, does the Bill still contemplate (or assume) that all elements of the conduct element would be apparent simply from examining the content? Does that differ as between the proactive duty and the reactive duty? For offences that apply only in part of the United Kingdom, does the Bill contemplate that the platform would have information about their applicability (or not) to a given item of content[16]?

4.3 It is notable that the Bill's definition of illegality refers at Clause 52(3)(b) to content that "when taken together with other regulated user-generated content present on the service, amounts to a relevant offence". That raises two further questions:

4.3.1 Is it contemplated that the platform would keep track of all content that might in the future amount to an offence in combination with other, as yet unknown, user content? If so, how is the platform to predict which content that would apply to?

4.3.2 Does this provision contemplate that *only* other content present on the service would be assessed, to the exclusion of extrinsic factual information?

That would be consistent with the Bill's apparent assumption that legality is capable of being assessed by a proactive automated filtering system, which by its nature operates on the basis of the information available on the platform. It would not, however, sit well with the reactive duty under Clause 9(3)(c), since a third party notifier would be likely to provide the platform with information not present on the platform.

In any event, for offences with a contextual aspect to the conduct element an assumption that legality could be assessed on the basis only of information present on the platform would not hold true. Context relevant to legality would include extrinsic factual matters.

Obvious examples of such offences would include the harassment offence and offences framed in terms of subjective effects on those who may encounter the content (such as the S.5 Public Order Act 1986 offence mentioned by the Human Rights Committee, S.4A of the Public Order Act 1986, the new

[16] Examples include the new harmful communications offence, which extends only to England and Wales (the existing S.127 Communications Act 2003 would remain in force in Scotland and Northern Ireland); the statutory codification of the public nuisance offence in the Police etc Act 2022, which extends only to England and Wales; the harassment offence under the Protection of Harassment Act 1997 (designated a priority offence), which extends to England, Wales and Northern Ireland; and sections 4, 4A and 5 of the Public Order Act 1986 (designated as priority offences), which extend to England and Wales. The illegality duties under the Bill extend to England, Wales, Scotland and Northern Ireland. Clause 52(9) provides that "For the purposes of determining whether content amounts to an offence, no account is to be taken of whether or not anything done in relation to the content takes place in any part of the United Kingdom." Does that have the effect, for the purposes of the illegality duties, of extending such offences across the whole of the UK?

harmful communications offence[17] and the newly codified public nuisance offence[18]).

4.4 If intention and available defences (such as reasonable excuse) are meant to form part of the platform's assessment of legality, those will inevitably bring into play extrinsic facts that are not apparent from merely examining the content in question.

4.5 These considerations reveal a triple bind:

4.5.1 If the Bill's approach to illegality excludes consideration of significant elements of a given offence (whether intention, available defences, or relevant context), the illegality duty will inevitably overreach and require removal of unexceptional[19] legal content.

4.5.2 If the Bill does not exclude those elements of the offence from consideration, but requires information to be taken into account that would not be available to the platform, then adjudging illegality is either rendered impossible or descends into arbitrary guesswork.

4.5.3 If the intent of the Bill is that (at least for proactive automated systems) the illegality duty should apply only to illegality that is capable of being adjudged merely by examining the user-generated content itself (either alone or in combination with other such content present on the platform), then it is unclear why in-scope offences (whether priority or non-designated) should include offences that depend significantly on intention, available defences (such as reasonable excuse), and factual context that cannot be apparent from the content itself.

5. ADJUDGING ILLEGALITY – TECHNICAL AND HUMAN PROCESSES

5.1 **What technical or human processes** does the Bill contemplate that a platform is capable of bringing to bear on the information contemplated to be available?

5.2 To restate the obvious, a platform cannot bring a technical or human process, however sophisticated, to bear on information that it does not have. Nor can it be a mind reader[20]. The implications of this may differ for proactive versus reactive duties.

---

[17] For an analysis of this offence and its interaction with the illegality duties under the draft Bill see Cyberleagle, *Licence to Chill* https://www.cyberleagle.com/2021/11/licence-to-chill.html.

[18] For an analysis of this offence see Cyberleagle, *Seriously annoying tweets* https://www.cyberleagle.com/2021/04/seriously-annoying-tweets.html.

[19] It might be suggested that there exists a penumbra of 'nearly illegal' content whose removal can be regarded with equanimity. If that proposition has any validity (which from a rule of law perspective of setting clear and precise boundaries is questionable), it must depend on the balance within any given offence between the conduct element, intention and available defences. For instance, the examples given by the Terrorism Legislation Reviewer illustrate that where an offence is founded on broadly drawn conduct and the heavy lifting of circumscribing the offence is performed by intention and available defences, a 'nearly illegal' penumbra would include swathes of unexceptional content.

[20] At best, and depending on the level of confidence required, it may be possible in some circumstances to infer intention from the contents of a post.

5.3 For **priority illegal content** the Bill contemplates proactive monitoring, detection and removal technology operating in real time or near-real time. There is no obvious possibility for such technology to inform itself of extrinsic information about a post, such as might give rise to a defence of reasonable excuse, or which might shed light on the intention of the poster, or provide relevant external context.

5.4 The Human Rights Committee observed in relation to S.5 Public Order Act 1986[21] ('threatening or abusive words or behaviour, or disorderly behaviour' that is likely to cause 'harassment, alarm or distress'):

> "It is hard to see how providers, and particularly automated responses, will be able to determine whether content on their services fall on the legal or illegal side of this definition."

5.5 The Human Rights Committee focused on the difficulty of distinguishing abusive from merely offensive, and the likelihood of causing someone 'distress'. Other aspects of the S.5 offence would also be difficult or impossible for a platform's systems to fathom, such as the requirement that the words be used "within the hearing or sight of a person"[22] likely to be caused harassment, alarm or distress thereby. The section also provides not only a defence of reasonableness, but also a defence that the defendant "had no reason to believe that there was any person within hearing or sight who was likely to be caused harassment, alarm or distress".

5.6 The S.4A offence (which like S.5 the Bill designates as a priority offence and so would be subject to proactive monitoring, detection and filtering duties) requires that someone is actually caused harassment, alarm or distress. The Law Commission observed[23]:

> "The High Court has described the words harassment, alarm and distress as being "relatively strong words" and clarified that "distress" requires "real emotional disturbance and upset"."

In the context of the proactive illegality duty, on what factual basis does the Bill contemplate that a platform's real time monitoring and filtering systems would adjudge whether someone was actually caused harassment, alarm or distress?

5.7 Even if (for instance) there were other posts responding to the post in question, is it contemplated that a platform's systems would be capable of determining the degree of upset merely by examining responsive posts (or at least such posts as the system

---

[21] The Bill designates S.5 as a priority offence.

[22] This provision gives rise to the question whether S.5 has any application at all online. In its Consultation Paper *Harmful Online Communications: The Criminal Offences* (11 September 2020) the Law Commission noted: "…it is arguable that the words "within the sight or hearing of a person likely to be caused harassment or distress" imply that the defendant must be present when that person views (or could have viewed) the offending material. It is not clear whether both the defendant and victim being online simultaneously would meet the requirement…". Would a provider have to decide whether the offence is capable of being committed online? Or does its inclusion in Schedule 7 of the Bill settle the matter? If so, would the platform still have to know whether poster and reader were online simultaneously?

[23] *Ibid.*

detected)? Is it contemplated that, when looking at responsive posts, the platform's systems would be able to distinguish upset from e.g. anger?

5.8 If there are no responsive posts (or none detected), on what factual basis are the platform's automated systems meant to determine whether harassment, alarm or distress has been caused?

5.9 For **non-designated illegal content** the platform is under a reactive duty to remove it swiftly "where the provider is alerted by a person to the presence of any illegal content, or becomes aware of it in any other way."[24] The platform's reactive removal duty is triggered only if and when the awareness threshold is reached. A person notifying the platform may choose to provide relevant extrinsic information which is not present on the platform.

5.10 Thus (assuming that 'illegal content' is interpreted so as to render extrinsic information relevant[25]), the Bill appears to contemplate that the platform could be put in possession of more information than would be possible within a proactive automated filtering environment. Of course that information might still be insufficient to fix the platform with awareness of illegality[26].

6. IMPLICATIONS FOR RULE OF LAW AND ECHR COMPATIBILITY

6.1 Legal certainty is a familiar concept from ECHR law, embodied in the 'prescribed by law' requirement. Aversion to vagueness (the legality principle) is also part of English domestic law. In the context of criminal law the objection to vagueness was spelt out by the House of Lords in *R v Rimmington*[27], citing the US case of *Grayned*:

> "Vagueness offends several important values … A vague law impermissibly delegates basic policy matters to policemen, judges and juries for resolution on an ad hoc and subjective basis, with the attendant dangers of arbitrary and discriminatory application."

6.2 The objection to vagueness is a domestic rule of law principle that applies to the law generally. Lord Diplock referred to it in a 1975 civil case (*Black-Clawson*):

---

[24] Assuming that this requires awareness of illegality, rather than merely awareness of a claim that certain content is illegal (and also assuming that intention and available defences are relevant to illegality), it bears some similarities to the circumstances in which a hosting provider may lose the protection of the liability shield provided by the ECommerce Directive (discussed here: Cyberleagle, *The Electronic Commerce Directive – a phantom demon?* https://www.cyberleagle.com/2018/04/the-electronic-commerce-directive.html). In that regard the observation of Eady J. in *Bunt v Tilley* (followed in subsequent caselaw) is noteworthy: "In order to be able to characterise something as "unlawful" a person would need to know something of the strength or weakness of available defences." (See further *Internet Law and Regulation*, G. Smith et al, Ed. 5 (Sweet and Maxwell) para 5-179).

[25] As to which, see Sections 3 and 4 above.

[26] The ECommerce Directive caselaw contains examples of notifications that were held insufficient to fix the host with knowledge of illegality (*Internet Law and Regulation*, paras 5-179 to 5-181).

[27] *R v Rimmington* [2005] UKHL 63.

> "The acceptance of the rule of law as a constitutional principle requires that a citizen, before committing himself to any course of action, should be able to know in advance what are the legal consequences that will flow from it."

6.3 The legal consequence that may flow for the user of a U2U platform is to have a post removed or otherwise interfered with as a result of the platform seeking to discharge its illegality safety duties[28].

6.4 As the government has acknowledged in its ECHR Memorandum, the Bill's illegality safety duties engage Art 10 ECHR[29]. Two especially relevant ways in which those duties could potentially be incompatible are:

6.4.1 **Prescribed by law**. The first step in an ECHR analysis is that the state's interference with a right has to be prescribed by law. This requires not just a legal framework, but that the legal basis should have the quality of law. The rule should be sufficiently clear and precise to enable the individual to predict in advance with reasonable certainty whether their speech will be interfered with (i.e. to foresee with reasonable certainty the consequences of their conduct). This is, in effect, a rule of law protection against arbitrariness.

6.4.2 **Proportionality**. In the context of the illegality duties, a disproportionate interference is most likely to occur through collateral damage to legitimate speech.

6.5 **Prescribed by law** It may be tempting to assume that because criminal offences are prescribed by law (assuming for the purposes of this analysis that each offence is sufficiently clear and precise), no prescribed by law issue can arise with a duty to enforce those offences. That, however, does not follow. If a nominally clear and precise rule is enforced by arbitrary means the necessary predictability of interference with the right will be lacking; all the more so if the arbitrariness does not consist in individual lapses, but is inherent in the enforcement framework.

6.6 The most obvious likely violation of the prescribed by law requirement would be if the illegality duties oblige platforms to adjudge illegality on the basis of information that is not, and cannot be, available to them (the interpretation at para 4.5.2 above, in relation to proactive monitoring). That inevitably results in arbitrary guesswork (at best) on the part of the platform and inability on the part of the user to predict with reasonable certainty what speech of theirs is liable to be interfered with by a platform seeking to carry out its safety duty.

6.7 What is the position if, instead, the duty requires the platform to assess illegality by reference only to those elements of illegality that are apparent from the user-generated content present on the system (which could arise under the interpretations at paras 4.5.1 or 4.5.3 above, in relation to proactive monitoring)? In that situation the platform is no longer determining illegality as defined in the various criminal offences. It is

[28] As a matter of ECHR law the Art 10 rights of potential readers of the posts are also engaged.

[29] Although the government Memorandum suggests that some of the speech affected might be excluded from Art 10 protection by virtue of Art 17 (abuse of rights), it does not argue that all affected content would be outside Art 10 protection, nor that any provisions of the Bill would only affect excluded content.

adjudging expanded versions of those criminal offences, shorn of limiting ingredients such as intention and available defences.

6.8 Such a rule would not necessarily breach the prescribed by law principle, but would do so if the expanded version of an offence were so broad as to be open to abuse through selective enforcement[30]. Additionally, inability to assess any still relevant extrinsic factual context would result in arbitrariness. Expanded versions of offences would also raise issues of necessity and proportionality, as the result of encompassing unexceptional content[31].

6.9 **Conclusion** This analysis may suggest that for a proactive monitoring duty founded on illegality to be capable of compliance with the 'prescribed by law' requirement, it should be limited to offences the commission of which can be adjudged on the face of the user content without recourse to further information.

6.10 Further, proportionality considerations may lead to the perhaps stricter conclusion that the illegality must be manifest on the face of the content without requiring the platform to make any independent assessment of the content in order to find it unlawful[32].

6.11 The government's ECHR Memorandum asserts compliance with the 'prescribed by law' requirement in the following terms:

> "The Bill establishes an overarching regulatory framework governing the treatment of content online. In addition to the detailed provisions contained in the primary legislation, its application in specific cases will be given a higher degree of legal certainty through the exercise of delegated powers by the Secretary of State and the issuing of codes of practice by OFCOM."

6.12 The Memorandum does not address the arbitrariness identified above in relation to proactive illegality duties, stemming from an obligation to adjudge illegality in the legislated or inevitable practical absence of material facts. Such a vacuum cannot be filled by delegated powers, by an Ofcom code of practice, or by stipulating that the platform's systems and processes must be proportionate.

6.13 **Prior restraint** Illegality duties, especially those involving real time or near-real time detection, filtering and take down of content, may also be viewed through the prism of prior restraint.

---

[30] Cf Gavin Millar QC Opinion for Index on Censorship, May 2022, para 36.

[31] The degree to which this would occur would vary from one offence to the next, depending on the constituent elements of each offence. The Terrorism Legislation Reviewer has illustrated its application to terrorism offences.

[32] *Poland v European Parliament and Council*, (Case C-401/19, 26 April 2022), a decision of the CJEU applying ECHR caselaw to platform filtering obligations in the context of copyright. It is also apparent that the CJEU regarded a requirement for complaint and redress mechanisms as insufficient on its own to ensure compatibility with the right of freedom of expression.

6.14 Prior restraint at the behest of the state can take two forms: pre-publication censorship, or interim restraint prior to full consideration of the merits. The latter is illustrated by the ECtHR decision in *Yildirim v Turkey*[33]:

> "The measure was to remain in place until such time as a decision was given on the merits or the illegal content of the site hosted by Google Sites was removed (section 9 of Law no. 5651). It therefore constituted a prior restraint *as it was imposed before a ruling had been given on the merits*." (emphasis added)

6.15 Mandated proactive detection and filtering in real time or near-real time is prior restraint of the first kind. Mandated reactive takedown triggered by notification falls somewhere between the two: it is not pre-publication censorship, but equally there is no mechanism for the merits to be considered later by an independent tribunal.

6.16 The validity of characterising at least proactive detection and filtering obligations as a form of prior restraint is illustrated by the recent CJEU decision in *Poland v The European Parliament and Council*[34], which applied *Yildirim* to those kinds of provisions in the context of copyright. The significance of *Yildirim* lies in the especially strict scrutiny that is applied to provisions imposing prior restraint:

> "The Court reiterates that Article 10 does not prohibit prior restraints on publication as such. … On the other hand, the dangers inherent in prior restraints are such that they call for the most careful scrutiny on the part of the Court." (*Yildirim*, para 47)

6.17 The presumption against prior restraint has a long history in the common law, and subsequently under the ECHR and the Human Rights Act 1998. Indeed, in its Consultation Paper on a Modern Bill of Rights the government proposes to strengthen the HRA's S.12 protections for freedom of expression and is considering raising the prior restraint threshold under S.12(3)[35].

6.18 The government's Human Rights Memorandum for the Online Safety Bill does not analyse the illegality duties from the perspective of prior restraint. Nor does it mention *Yildirim*.

6.19 **Collateral damage** Excessive collateral damage to legitimate speech is disqualifying on proportionality grounds[36]. It is difficult to see how a regime with built-in overreach could be cured by requiring an inherently disproportionate regime to be operated proportionately, whether by Ofcom, by service providers, or by both in combination.

---

[33] *Yildirim v Turkey*, ECtHR, App No. 3111/10, 18 Dec 2012 at [52].

[34] Above, n. 32.

[35] Human Rights Act Reform: A Modern Bill Of Rights (Dec 2021) paras 213 – 214.

[36] English caselaw on intellectual property site blocking orders, which has applied an ECHR proportionality analysis, has adopted a proportionality standard of insubstantial or *de minimis* collateral interference with legal content. *Twentieth Century Fox Film Corp & Ors v British Telecommunications Plc* [2011] EWHC 1981 (Ch) (186, 201) and subsequent cases. The CJEU in *Poland v The European Parliament and Council* appears to have set a proportionality standard of no collateral damage to legal content, subject only to the possibility of correcting errors that may occur.

6.20 **Freedom of expression duty** The Bill requires a platform to have regard to the importance of protecting users' right to freedom of expression within the law. Since that duty is predicated on the legality of the user's speech, it takes no further the question of how a platform should perform the task of adjudging legality and illegality.

6.21 **Summary** Depending on its interpretation the Bill appears either:

6.21.1 to exclude from consideration essential ingredients of the relevant criminal offences, thereby broadening the offences to the point of arbitrariness and/or disproportionate interference with legitimate content; or

6.21.2 to require arbitrary assumptions to be made about those essential ingredients, with similar consequences for legitimate content; or

6.21.3 to require the existence of those ingredients to be adjudged, in circumstances where extrinsic factual material pertaining to those ingredients cannot be available to a filtering system.

In each case the result is arbitrariness (or impossibility), significant collateral damage to legal content, or both. It is not easy to see how on any of those interpretations the Clause 9(3) proactive filtering obligation could comply with either the prescribed by law requirement or the proportionality requirement.

7. OTHER MATTERS

7.1 This submission has focused on the illegality duties contained in Clause 9(3). A separate illegality duty in Clause 9(2) relates to mitigating and managing the risk of harm to individuals identified in the most recent illegal content risk assessment.

7.2 Presumably the intention, although the clause does not say so, is that this duty should apply only to risk of harm *caused by* in-scope illegal content. There appears to be a drafting issue. Assuming that the intention is to limit the duty in that way, the ECHR compatibility issues identified above around the section 9(3) duties will also apply to these duties.

7.3 If the S.9(2) duties are not intended to be limited to harm *caused by* illegal content, that raises significant issues of scope, since the duties would then appear to extend to any kind of harm (as defined by Clause 187), whether illegal or not, so long as it was 'identified' in the most recent illegal content risk assessment.

7.4 As well as raising further questions of legal certainty and proportionality, that would be inconsistent with the government's assumed intention of addressing legal but harmful content only by means of separate duties elsewhere in the Bill.