IN THE MATTER OF:

THE PRIOR RESTRAINT PROVISIONS IN THE ONLINE SAFETY BILL

ADVICE

I. INTRODUCTION AND SUMMARY OF ADVICE

1. We are asked to advise the Open Rights Group on whether certain provisions in the Online Safety Bill ("the Bill") relating to the prior restraint of content published online are lawful.

2. Clause 9(2)(a) of the Bill places a duty on online platforms, such as Facebook and Twitter, to prevent users from "*encountering*" certain "*illegal content*", which includes a wide range of offences. While AI or other algorithmic screening of content is not expressly mandated, the Bill operates on the assumption that regulated providers will use such technology in order to comply with the illegal content provisions.

3. A key concern about this duty is that it effectively amounts to prior restraint, as it will require the interception and blocking of online communication (meaning that the content is never posted). Strikingly, there is no requirement in the Bill for individuals to be notified that their content has been blocked, let alone provided with reasons for the blocking. We understand that the Open Rights Group is concerned that automatic programmes are prone to error, especially in interpreting contextual meaning, and are unable to perform the textual and legal analysis needed to distinguish lawful and unlawful content. In particular, they have been shown to entrench and perpetuate bias against minority groups by, for example, disproportionately and incorrectly identifying language and actions of such groups as harmful or dangerous.

4. The current version of the Bill (dated 18 January 2023) is under consideration at the Committee stage of the House of Lords beginning on 22 June 2023.

5. We are asked:

    a. Whether there are adverse or unintended consequences that might result from the passage into law of Clause 9(2)(a); and
    b. Whether Clause 9(2)(a) is compatible with the European Convention on Human Rights ("the Convention").

6. In our view, the Bill, if enacted in its current form, will represent a sea change in the way public communication and debate are regulated in this country. It risks fundamental encroachments into the rights of freedom of expression to impart and receive information. The Bill will require social media platforms, through AI or other software, to screen material before it is uploaded and to block content that could be illegal. That is prior restraint on freedom of expression and it will occur through the use by private companies of proprietorial, and no doubt secret, processes. It will give rise to interference with freedom of expression in ways that are unforeseeable and unpredictable and through entirely opaque processes, and in ways which risk discriminating against minority religious or racial groups.

7. In summary, our advice is that the legal consequences of that are:

    a. Accordance with the law. In our view there is a significant risk that the prior restraint provisions will lead to interference with freedom of expression which is not "*prescribed by law*" for the purposes of Article 10 of the Convention. We are concerned that users will not be able to foresee and predict in advance which content will be subject to prior restraint. That is so given (i) the vast range of broadly defined potential illegal content; and (ii) the lack of transparency and accountability relating to the processes that will be used by regulated providers to determine when conduct is potentially unlawful. In addition there do not appear to us to be adequate safeguards against arbitrariness. In particular, the complaints process available to users is likely to be ineffective and users may not even know their content has been blocked (let alone provided with reasons for it).

b. Right to freedom of expression. We consider that, taken with the rest of the Bill in its current form, Clause 9(2)(a) also gives rise to a risk of a disproportionate interference with the Article 10 rights of users to post and view content. The Bill heavily incentivises regulated providers to block content that might be illegal by a robust sanctions regime, which includes very significant fines. In contrast, the redress available for users who have had posts blocked is extremely limited. There is a real risk service providers will be overly incentivised to block content and that a significant number of lawful posts will be censored, without any genuine effective redress for individuals.

c. Discrimination. Adverse and unintended consequences are likely to include that the prior restraint provisions disproportionately and unjustifiably impact certain ethnic or other monitory groups. That is of particular concern as automated discriminatory decision making will take place in an AI "*black box*" such that it is simply not possible to scrutinise to determine how prior restraint is operating and whether it is doing so in a discriminatory manner. There is a significant and growing body of research that suggests that automated technology frequently contains inherent biases. There is a clear risk that the screening systems will disproportionately block content relating to or posted by minority ethnic or religious groups. This is particularly concerning because the systems will be operated by private providers who are not subject to the Public Sector Equality Duty ("PSED") and users have very limited prospects of challenging individual decisions as being discriminatory. This could give rise to discriminatory interference with rights to receive and impart information which would be contrary to Article 14 of the Convention, taken with Article 10.

8. We are concerned that this is an overreaching Bill that risks capturing a range of legitimate content that is important to public debate and preventing it being published. As presently formulated it is liable to have a serious chilling effect on free speech online and we consider it goes beyond what is necessary to meet the purported aims of the Bill.

II. OVERVIEW OF RELEVANT PROVISIONS

9. This opinion is concerned primarily with the provisions relating to "prior constraint". The relevant provisions are as follows.

Regulated services/ providers

10. Clause 2(1) defines a "*user to user service*" as "*an internet service by means of which content that is generated directly on the service by a user of the service, or uploaded to or shared on the service by a user of the service, may be encountered by another user, or other users, of the service*".

11. Clause 3 defines regulated services, also known as "*Part 3 services*". A "*user to user service*" is regulated if it: (a) has links with the UK; and (b) is not listed as exempt in Schedule 1 or 2 (Clause 3(2)). A service has links with the UK if: (a) the service has a significant number of users in the UK; or (b) UK users form one of the target markets for the service (Clause 3(5)). Further, Clause 3(6) says that a service has links with the UK if: (a) it is capable of being used by individuals in the UK; and (b) there are reasonable grounds to believe that there is a material risk of significant harm to individuals in the UK by user generated content or search content.

12. The prior restraint duty is therefore likely to be imposed on social media platforms, such as Facebook, Instagram, TikTok and Twitter, with users of those platforms potentially having their content blocked.

Duties on regulated service providers

13. Part 3 imposes duties of care on providers of regulated user-to-user services. Clauses 6 and 7 detail outlines which duties apply to user to user services, and their scope.

Prior restraint obligations

14. Clause 9 provides for safety duties about illegal content. In particular, Clause 9(2) imposes a duty "*to take or use proportionate measures relating to the design or operation of the service to –*

    (a) Prevent individuals from encountering priority illegal content by means of the service,
    (b) Effectively mitigate and manage the risk of the service being used for the commission or facilitation of a priority offence, as identified in the most recent illegal content risk assessment of the service, and
    (c) *Effectively mitigate and manage the risks of harm to individuals, as identified in the most recent illegal content risk assessment of the service*" (emphasis added).

15. The use of the terms "*prevent*" and "*encountering*" indicate to us that regulated providers will, in practice, be required to screen content at the point of upload to the relevant platform using AI algorithms or other software. Indeed it is difficult to see how the duty could be fulfilled without doing so. If the content is deemed to be priority unlawful content, the upload will be blocked by the screening technology.

16. Clause 9(7) places a duty on providers to include information in the terms of service about the use of "*proactive technology*" used for the purpose of complying with the duty in section 9(2). Clause 202 defines "*proactive technology*" as including "*content moderation technology*". Clause 202(2) defines content moderation technology as "*algorithms, keyword matching image matching or image classification – which (a) analyses relevant content to assess whether it is illegal content…*".

17. "*Illegal content*" means content that amounts to a relevant offence (Clause 53(2)). Clause 53(3) provides that "*Content*", which is "*words, images, speech or sounds*", amounts to a relevant offence if:

    a. the use of the words, images, speech or sounds amounts to a relevant offence,
    b. the possession, viewing or accessing of the content constitutes a relevant offence, or

c. the publication or dissemination of the content constitutes a relevant offence.

18. "*Priority illegal content*" is defined in Clause 53(7)-(10) as: (a) terrorism content (i.e. content that amounts to an offence specified in Schedule 5 (Clause 53(8)); (b) content relating to child sexual exploitation and abuse (i.e. content that amounts to an offence specified in Schedule 6); and (c) content that amounts to an offence specified in Schedule 7 (which includes assisting suicide, fraud and public order offences).

19. The range of terrorism offences specified in Schedule 5 is broad. It includes, for example, section 12(1A) of the Terrorism Act 2000: expressing an opinion or belief supportive of a proscribed organisation. It also covers inchoate terrorist offences, which includes aiding, abetting, counselling or procuring the commission of a specified terrorist offence (paragraph 4).

20. Schedule 6 sets out the child sexual exploitation and abuse offences. Schedule 7 details other priority offences, ranging from assisting suicide to contraventions of financial services restrictions. Of particular note are the following offences:

    a. Use of words likely to cause harassment, alarm or distress (section 5 of the Public Order Act 1986);
    b. Use of words or behaviour or display of written material that is abusive or insulting and which is likely to stir up racial hatred (section 18 of the Public Order Act 1986);
    c. Assisting unlawful immigration (section 25 of the Immigration Act 1971);
    d. Fraud by false representation (section 2 of the Fraud Act 2006); and
    e. The offence of aiding, abetting, counselling or procuring the commission of an offence specified in the Schedule.

21. Clause 170 provides that, in determining whether content is illegal, providers should have "*reasonable grounds to infer*" that the content is illegal. A provider may have reasonable grounds if a provider (Clause 170(6)):

a. Has reasonable grounds to infer that all elements necessary for the commission of the offence, including the mental elements, are present or satisfied; and
b. Does not have reasonable grounds to infer that a defence to the offence may successfully be relied on.

Consequences of non-compliance with Clause 9 duty

22. There is a sanctions regime in Part 7 of the Bill which seeks to ensure compliance with the safety duties, including the duty in Clause 9(2)(a).

23. Under Clause 118, OFCOM may give a provisional notice of contravention to a regulated provider if there are reasonable grounds to believe it is failing to comply with an enforceable requirement (which includes the Clause 9 duties). If, following receipt of representations from the regulated provider, OFCOM is satisfied that the provider has failed to comply with a notified requirement, it may make a "*confirmation decision*" (Clause 120). This may require the provider to take steps (Clause 121), or to take steps to use a kind of proactive technology (Clause 124) and/ or pay a penalty (Clause 125). The penalties include fines of up to £18 million or 10% of the provider's global revenue (Schedule 13, paragraph 4(1)).

Provisions protective of users' rights

24. By contrast to the draconian enforcement conditions to encourage prior restraint, the provisions which enable users to challenge over-zealous action, which blocks lawful content, taken pursuant to Clause 9(2)(a) are limited. Regulated providers have a duty to operate a complaints procedure that allows for "*appropriate action*" to be taken in relation to "*relevant kinds of complaint*" (Clause 17(2)). This enables users to complain about the use of proactive technology that results in content being restricted (Clause 17(4)(e)). However, there is nothing in Clause 17 about the timescales in which a complaint must be addressed. There do not appear to be any enforcement processes for a failure adequately to address complaints and no entitlement to compensation.

Freedom of expression

25. Clause 18 imposes on regulated providers a duty (including in relation to the Clause 9 safety duty) to "*have particular regard to the importance of protecting users' right to freedom of expression within the law*" (Clause 18(2)). No further detail is provided however. Furthermore, the duty goes no further than requiring "*regard*" to be had to the importance of freedom of expression, and provides for no consequence if even that attenuated duty is breached.

26. Clause 13(2) imposes a duty on service providers to operate a service designed to "*ensure that the importance of the free expression of content of democratic importance is taken into account*" when making decisions about removing content. "*Content of democratic importance*" is limited in Clause 13(6) to news publisher or regulated user-generated content that is "*specifically intended to contribute to democratic political debate*" in the UK. This is a narrow duty that goes no further than requiring free expression to be "*taken into account*".

27. Clause 14 imposes a duty on service providers to take certain steps before taking action in relation to "*news publisher content*". Those steps include notifying the news publisher, considering representations and providing reasons for the ultimate decision (Clause 14(3)). This duty provides a limited number of safeguards but only in respect of news publishers.

28. Clause 15 imposes duties on service providers to protect "*journalistic content*", which is defined as (a) news publisher or regulated user-generated content (b) that is generated for the purposes of journalism and (c) that is UK linked. This, again, is a narrow duty that only requires freedom of expression to be "*taken into account*" (Clause 15(2)).

29. Clause 143 requires OFCOM to make an annual statement on how they ensure that their online safety functions have been exercised compatibly with Article 10 of the Convention.

Code of Practice

30. OFCOM must prepare and issue a code of practice for providers of regulated services describing measures recommended to comply with the duties in Clause 9 (Clause 36(1)). At this stage we do not have detail about what the code of practice will cover. Clause 44(2)(a) states that regulated providers will be treated as complying with Clause 18 if "*the provider takes or uses such of the relevant recommended measures as incorporate safeguards to protect users' right to freedom of expression within the law*".

III. RIGHT TO FREEDOM OF EXPRESSION: OVERVIEW

General principles

31. Article 10(1) of the Convention provides that everyone shall have the right to freedom of expression, which includes "*freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers*". Article 10(2) states:

> "*The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary*" (emphasis added).

32. Freedom of expression constitutes one of the essential foundations of democratic society (*Handyside v United Kingdom* (1979-80) 1 EHRR 737 at [49]). It applies not only to information or ideas that are favourably received but also to those that offend, shock or disturb [59]. The exceptions to freedom of expression must be construed strictly and the need for any restrictions must be established convincingly (*Stoll v Switzerland* (2008) 47 EHRR 59 at [101]).

33. The ECtHR has recognised the importance of information sharing on the internet, which "*plays an important role in enhancing the public's access to news and facilitating the dissemination of information in general*" (*Times Newspapers Ltd v United Kingdom (nos 1*

and 2) [2009] EMLR 14 at [27]). Article 10 protects both the content of information and the means by which it is communicated (*Ahmet Yildrim v Turkey* (2012) App No 3111/10 at [50]). Digital and online communication is an essential exercise of the right to freedom of expression (*Ahmet Yildrim v Turkey* at [48]-[49]; *Vladimir Kharitonov v Russia* (2020) App No 10795/14 at [33]).

34. Adverse consequences for users or platforms in allowing content to be published online could have a "*chilling effect*" on freedom of expression (*Weber & Saravia v Germany* (2008) 46 EHRR SE5 at [78]; *Brasilier v France* (App No 71343/01) 2006 at [43]). As set out below (at [63]), prior restraint, which prevents material ever being published, is regarded as a particular significant interference with Article 10 rights and thus require particularly compelling justification.

Interference with Article 10

35. Interferences with the right to freedom of expression may occur in the following ways in this context.

36. First, the rights of users of regulated platforms may be engaged. Content being blocked by means of prior restraint amounts to a straightforward interference with the right to express and receive that information. That may happen wrongfully if the AI or other algorithm misidentifies legitimate content as priority illegal content. A corollary of this is that individual users may be deterred from communicating due to fears of having their content screened and potentially blocked, having a chilling effect on free speech.

37. Second, the rights of the regulated platforms themselves may be engaged. In *Tamiz v United Kingdom* [2018] EMLR 6, the ECtHR held that imposing liability for and/or compelling the removal of content by provider of a blogging platform engaged the platform's Article 10 rights [6].

IV. ARTICLE 10: PRESCRIBED BY LAW

Key legal principles

38. For an interference with freedom of expression to be lawful it must be "*prescribed by law*."

39. The European Court of Human Rights has considered what is required for a measure to be adequately prescribed by law. It has held that the measure in question must (a) have "*some basis in domestic law*" and (b) must be "*compatible with the rule of law*", which means that it should comply with the twin requirements of "*accessibility*" and "*foreseeability*" (*Sunday Times v United Kingdom* (1979) 2 EHRR 245; *Sliver v United Kingdom* (1983) 5 EHRR 347; and *Malone v United Kingdom* (1984) 7 EHRR 14). In addition, the law "*must afford adequate legal protection against arbitrariness and accordingly indicate with sufficient clarity the scope of discretion conferred on the competent authorities and the manner of its exercise*" (*S v United Kingdom* (2009) 48 EHRR 50 at [95] and [99]).

40. The Strasbourg approach has been considered and applied in the domestic courts which have reiterated that an interference will not be in accordance with the law unless it has some basis in domestic law, is accessible and foreseeable (*In re Gallagher* [2020] AC 185 at [16]-[23]). Lord Sumption in *Gallagher* defined the foreseeability requirement as meaning that it must be possible for a person to foresee the consequences of their actions for them. An individual "*must be able – if needs be with appropriate advice – to foresee, to a degree that is reasonable in the circumstances, the consequences which a given action may entail*" (*Sunday Times v United Kingdom* [49] cited in *Gallagher* [19]). In addition the law should not "*confer a discretion so broad that its scope is in practice dependent on the will of those who apply it, rather than on the law itself*" (*Gallagher* [17]), and there must be "*safeguards which have the effect of enabling the proportionality of the interference to be adequately examined*" (*R (T) v Chief Constable of Greater Manchester and others* [2015] AC 49 at [144]). The rules governing the scope and application of measures need not be statutory, provided that they operate within a framework of law and that there is an effective means of enforcing them (*R (Catt) v Association of Chief Police Officers* [2015] AC 1065 at [11]).

41. As the domestic and Strasbourg authorities have made clear, in order for a measure that interferes with fundamental rights to be "*prescribed by law*" there must be sufficient protections against arbitrary interference. A key feature of such protection is the ability of individuals meaningfully to challenge, and for the courts to be able to review, the

unlawful exercise of a power that interferes with fundamental rights. That was considered by the Supreme Court in *R (Roberts) v Commissioner of Police of the Metropolis* [2016] 1 WLR 210 and in *Christian Institute v Scottish Mins* [2016] UKSC 61. *Roberts* concerned a police power to stop and search without prior suspicion. The Supreme Court held that the power was lawful because there was a clear regime governing the exercise of the power and because there were provisions which enabled the unlawful exercise of the power to be challenged and which thus provided sufficient protection against their arbitrary and disproportionate exercise. As the Supreme Court explained in *Christian Institute* at [81], when considering *Roberts:*

> *[Published policies] regulated the authorisation of stop and search, the [overall] operation and also the individual encounter between a police officer and a member of the public on the street. In relation to the exercise on the street of the stop and search power it not only gave officers detailed instructions, which were designed to ensure their proportionate use of such power, but also required them to explain to the individual who was to be searched the reason for the search, to record that reason in writing and make available to the affected individual a copy of that written record. … These provided adequate safeguards to enable the courts to examine the proportionality of any interference with fundamental rights.*

<u>Application of principles to the Bill</u>

42. Our understanding of how the Bill works and will apply in practice is as follows:

    a. Clause 9(2)(a) obliges regulated providers to screen information before it is shared on online platforms to ascertain whether it is priority illegal content.
    b. It is anticipated that regulated providers will discharge that obligation through AI or other algorithms that will screen the information before it is uploaded.
    c. This will likely be using proprietary algorithms that is operated by private companies and is not open to meaningful scrutiny.
    d. How such algorithms determine what amounts to priority illegal content will be a matter for the private providers (subject to any requirements in the code of practice).

e. There is no requirement in the Bill for providers to notify the user that the content has been blocked or to provide reasons for why it was blocked.
f. The most a user can do if he or she realises that content has been blocked is to make a complaint to the provider, but there are no specific timescales in which complaints must be considered or recourse / compensation for wrongful blocking.

43. In our view, this raises serious concerns about whether the prior restraint provisions are sufficiently "*prescribed by law*" to be consistent with Article 10.

Foreseeability

44. First and foremost, these provisions are likely not to be in accordance with the law (and therefore contravene Article 10) because it will not be possible, in practice, to foresee whether certain content will be blocked by service providers. That is because it will be determined by an AI or other system what amounts to "*priority illegal content*" and individuals will often not be able to predict or foresee what about their content might result in it being blocked or indeed understand why content has been blocked (if they find out it has been blocked at all).

45. We set out below a number of examples of how the Bill, if it became law, is likely to operate. For each there is a real risk that:

a. Innocent and important content will be wrongly blocked as "*illegal*" and it will be impossible to know on what basis that has occurred or to predict its occurrence;
b. In particular it will be extremely difficult, if not impossible, for an AI algorithm to assess whether the posting of content amounts to an offence and therefore extremely difficult if not impossible to predict what will lead to material being blocked; and
c. There will be a consequent chilling effect on free speech.

46. Public order offences. Section 5(1) of the Public Order Act 1986 provides, inter alia, that a person is guilty of an offence is he uses "*threatening or abusive words or disorderly*

behaviour". One of the defences a person may rely on is that his conduct was "*reasonable*" (section 5(3)(c)). "*Disorderly*" behaviour online could capture an extremely wide range of conduct. It is unclear how providers will determine what words are "*threatening or abusive"* or when words are abusive rather than merely insulting given that such determination requires an understanding of meaning in context. It is also extremely difficult to see how an AI algorithm could assess whether the content was reasonable or to predict how the algorithm will carry out such an assessment. It is likely that for section 5 (and other) offences, possible defences may simply be ignored by the algorithm because there is not sufficient information available from the content of the post itself.

47. Section 18 of the Public Order Act 1986 provides that "*(1) A person who uses threatening, abusive or insulting words or behaviour, or displays any written material which is threatening, abusive or insulting, is guilty of an offence if – (a) he intends thereby to stir up racial hatred, or (b) having regard to all the circumstances racial hatred is likely to be stirred up thereby*". The difficulty with predicting how an algorithm will determine if this offence has been committed is obvious. It is capable of capturing and blocking content that is considered merely "*insulting*" simply because it relates to a racial issue or a racial group. It is difficult to see how an algorithm could accurately determine the individual's intention and/or the relevant wider circumstances or how it will be possible to predict, or understand, how that will be determined.

48. The ECtHR has repeatedly observed that public order offences are broad and may be vague, but that is ameliorated by an individual's ability to seek legal advice (see e.g. *Kudrevicius and others v Lithuania* (2016) 62 EHRR 34 at [108]-[109], [113]). However, here there is no option to seek legal advice and the individual may have no way of knowing why their content was blocked or even that it was blocked because it was regarded as "*abusive"* or "*threatening"*.

49. Terrorism offences. Schedule 5 provides that various terrorist offences are priority illegal content. This includes, for example section 12 of the Terrorism Act 2000, which prohibits inviting support for proscribed organisations and expressing an opinion that is supportive of a proscribed organisation. It is difficult to predict how that will be determined by an algorithm. For example, Hamas is a proscribed organisation in the

UK. It is also in de facto control of Gaza. It is difficult to see how an AI or other algorithm will determine whether discussion of the political situation in Gaza constitutes legitimate debate or expressing an opinion supportive of Hamas, and there is a clear risk that lawful content will be blocked. That risks stifling debate about important political and geographical issues and doing so in ways that are unpredictable and unforeseeable.

50. Immigration offences. Assisting unlawful immigration under section 25 of the Immigration Act 1971 is listed in Schedule 7. The government has expressly indicated that this provision may be used to prevent people from posting photos of refugees arriving in the UK by boat if they "*show that activity in a positive light*"[1]. It is difficult to predict how an AI algorithm will decide what constitutes showing such an image in a "*positive light*". What if the image was posted without any text? Again, that may have a chilling effect on a debate of public importance relating to the treatment of refugees and illegal immigration.

51. Fraudulent offences. Fraud by false representation under section 2 of the Fraud Act 2006 is listed in Schedule 7. The mens rea of this offence requires the individual be "*dishonest*" about making a false representation, with the intention of making a gain for himself or a loss to another. It is very difficult to see how an AI algorithm can accurately ascertain whether an individual is being dishonest in posting content on a social media platform or what their intention is. That requires further investigation and it is unforeseeable, in those circumstances, what kind of content might be blocked. It is possible that an AI algorithm might block any content that it considers to be false as a precaution, irrespective of the individual's mens rea.

52. This uncertainty is reinforced by the inclusion of inchoate offences at paragraph 33 of Schedule 7, which includes (i) attempting or conspiring to commit one of the specified offences and (ii) aiding, abetting, counselling or procuring the commission of one of the offences. In each of the above scenarios, the relevant content is even more likely to be caught in unpredictable ways by the inchoate versions of the specified offences. Moreover, accessory liability (aiding and abetting) is dependent on the commission of

[1] Written statements - Written questions, answers and statements - UK Parliament

the primary offence by another party. It is difficult to see how an automated screening system will attempt to find out whether the primary offence took place.

53. The fundamental difficulty with the approach of the Bill is that, in many cases, it will not be possible to determine in advance whether content will be blocked as illegal. The criminal law operates with evidential and procedural protections that ensure all relevant circumstances can be considered carefully, including the perspective of the individual defendant. Individuals are able to obtain advice in order to enable them to determine, in advance, whether particular actions or particular statements are likely to be unlawful and restraints on freedom of expression will almost always occur *after* the expression has occurred through the operation of the criminal law. The Bill operates entirely differently. It will give rise to prior restraint on publication through automated processes that will determine potential illegality *in advance*. That will occur through private companies operating what will, no doubt, be unpublished and secret processes to detect potentially unlawful content. The risk that will operate in ways that are unpredictable, uncertain and impossible to properly regulate is obvious. These concerns are reinforced by the fact that different private companies will take different approaches, and thus something that is permitted on one platform may well be prohibited on another platform. We consider the result is likely to be significant interference with freedom of expression that is unforeseeable and which is thus not prescribed by law.

Inadequate safeguards against arbitrary restraint

54. In addition, the safeguards against arbitrariness do not appear to us to be adequate.

55. First, the complaints process is very likely to be ineffective (see below at [67]). Unlike in the regime considered, for example, in *Roberts*, there is no adequate process or remedy if there is a breach of an individual's Article 10 rights. The Bill envisages a complaints procedure, but there are no timescales and no right to compensation. There is no effective appeal process or right to an appeal in court.

56. Second, the Bill does not specify how regulated providers should approach developing or purchasing the screening algorithms: it is effectively up to the providers. Private

companies are not subject to the obligations under the Convention, meaning that there may be no potential claim under the HRA for a breach of their rights.

57. Third, there is no obligation for providers to provide reasons why content is not permitted to be uploaded or permit access to the algorithms that make such decisions, or even to notify them that the content has been blocked (c.f. *Roberts*). This prevents any meaningful scrutiny of the approach of the AI or other algorithms and prevents individuals from identifying flaws in the decision making.

58. Fourth, the provider need only have "*reasonable grounds*" for concluding that the content is illegal. That is a low threshold in circumstances where it results in the censoring of content. Whether conduct is illegal is a binary choice and there is no scope for providers to decide that content is partly illegal or that further investigation is required.

59. Finally, while there is a reference to safeguards and recommended measures that will be included in OFCOM's code of practice (see Clauses 18 and 44), we cannot at this stage be sure what they might include or whether they are legally sufficient. While codes of practice can in principle meet the "*prescribed by law*" requirement, we are surprised by how much is being left to the planned code, rather than being included in the Bill and subjected to proper Parliamentary scrutiny at this stage.

V. ARTICLE 10: PROPORTIONALITY / NECESSARY IN A DEMOCRATIC SOCIETY

60. The legitimate aims for interference with the right to freedom of expression are outlined in Article 10(2). Detecting and preventing the distribution of child sexual abuse, terrorist and other material online is, no doubt, a legitimate aim (protecting public safety and / or the rights of others). We consider, however, that the Bill is likely to give rise to a disproportionate interference with the Article 10 rights of users in pursuit of those legitimate aims.

Key legal principles

61. As explained by the ECtHR in *Handyside v United Kingdom*, in order to determine whether an interference is necessary in a democratic society, a court needs to examine the interference complained of in light of the case as a whole and determine whether the interference was "*proportionate to the legitimate aim pursued*" and whether the reasons adduced by the national authorities to justify it are "*relevant and sufficient*" [49]-[50].

62. In order for a measure to be proportionate and necessary in a democratic society, there must be no other means of achieving the same end that would interfere less seriously with the fundamental right concerned (*Glor v Switzerland* (App No 13444/04) 2009 at [94]).

63. Whilst the ECtHR has stated that, generally speaking, Article 10 does not prohibit all prior restraint (such as in the form of interlocutory and permanent injunctions), the dangers inherent in prior restraints call for the most careful scrutiny by the court (*Observer and Guardian v United Kingdom* (1992) 14 EHRR 153 at [60]). That is especially so in relation to the sharing of news, which is a "*perishable commodity and to delay its publication, even for a short period, may well deprive it of all its value and interest*" (*Observer and Guardian v United Kingdom* at [60]; *Cumpana and Mazare v Romania* (2005) 41 EHRR 14 at [118]). In *Ahmet Yildrim v Turkey*, the ECtHR found that such restraints must therefore form part of a legal framework ensuring tight control over the scope of any prior restraint and effective judicial review to prevent any abuses [64]. It is necessary to weigh up the various interests at stake, including the "*significant collateral effect*" of substantially restricting rights to receive information [66].

Analysis and application of principles

64. In our view, there is a clear risk that the prior restraint provision in the Bill will give rise to a disproportionate interference with Article 10 rights.

65. First, as detailed above, regulated providers, along with individuals, are likely to struggle to know whether certain posts are priority illegal content. There may be much content that comes close to the dividing line. The service provider need only have

"*reasonable grounds*" to "*infer*" that the content is illegal and there is a risk that a significant amount of non-illegal content will thus be blocked.

66. Second, there is a strong sanctions regime designed to punish regulated providers for failing to take action under Clause 9(2)(a) to prevent priority illegal content from being uploaded (see above at [23]). This includes confirmation decisions which mandate that the providers take certain steps. Regulated providers may be fined vast sums of money, of up to £18 million or 10% of the provider's global revenue (Schedule 13, paragraph 4(1)). This is clearly designed to ensure that providers take a robust approach to proactive restraint.

67. Third, in stark contrast, the consequences of erroneously blocking legal content are almost non-existent for regulated providers. Regulated providers are required to have a complaints mechanism, but:

    a. There are no specific sanctions on the regulated providers for non-compliance.
    b. There are no timescales that require complaints to be considered quickly. This is particularly important in this context where certain content is a "*perishable commodity*" and may be of little or no value if it is only uploaded weeks later after a lengthy appeal process.
    c. Users are not entitled to compensation for wrong decisions.
    d. There is no appeal process following a complaint.

68. Fourth, in light of the above, regulated providers are likely to be heavily incentivised to rely on overly cautious algorithms, that err on the side of ensuring that the risk of uploading priority illegal content is minimised as far as possible. This is even more so given the generous language of having "*reasonable grounds*" to believe content is illegal. This means that in practice service providers are likely to use algorithms which will capture much content that is not illegal. That is not likely to meet the "*necessary in a democratic society*" test, because it would be possible to adopt a system that interferes less with users' Article 10 rights (i.e. by rebalancing the incentives or setting a threshold that is higher than "*reasonable grounds*").

69. Fifth, the opportunity for monitoring the approach taken by regulated providers is likely to be extremely limited. AI or other software will be operated by private providers and are likely to be difficult if not impossible to analyse. There is currently no obligation for those algorithms to be monitored or analysed by a public body and it is very difficult to see how it will be possible to ensure they are operating in a proportionate and restrained way which gives sufficient weight to the importance of protecting freedom of expression.

70. Sixth, individuals will not be able to bring claims under the HRA 1998 against the regulated providers because they are not public authorities. It is difficult to see how users will have access to a judicial remedy to vindicate their Article 10 rights.

VI. DISCRIMINATION

71. We understand that the Open Rights Group is concerned that the prior restraint provisions may discriminate against certain minority groups. In short, bias and discrimination may mean that AI algorithms or other software disproportionately block content relating to or posted by minority ethnic groups. This would be contrary to Article 14 of the Convention, taken with Article 10.

Key legal principles

72. Article 14 of the Convention provides that:

> "*The enjoyment of the rights and freedoms set forth in the Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.*"

73. There are four limbs for the engagement of Article 14 (see e.g. *McLaughlin v SSWP* [2018] UKSC 48 at [15] per Baroness Hale). They are (a) whether the circumstances "*fall within the ambit*" of one or more of the Convention rights; (b) whether there been a difference of treatment between two persons who are in an analogous situation; (c) whether that difference of treatment was on the ground of one of the characteristics

listed or "*other status*"; and (d) whether there is an objective justification for that difference in treatment.

74. In relation to indirect discrimination, the second and third stages are replaced by the question: does the measure have a differential effect on persons with a protected characteristic (*R (SG) v Secretary of State for Work and Pensions* [2015] 1 WLR 1449 at [13]-[14])?

Analysis

75. We understand from our instructions that there are particular concerns in relation to the use of automated decision-making operating in manners that discriminate on grounds such as race or religion. There is a growing body evidence of AI and other automated systems disproportionately discriminating against particular minority groups. For example, AI systems have been shown to have racial dialect bias, where they "*systematically classify content aligned with the African American English (AAE) dialect as harmful at a higher rate than content aligned with White English (WE)*".[2] A hiring algorithm was found to disproportionately prefer men over women,[3] and chatbots have been found to rapidly learn and use racist language.[4] Concerns have been raised by partners of the Open Rights Group that the Bill may operate in ways that reinforce or amplify existing racism and injustice. For example: (i) an algorithm may identify specific language patterns on the basis that they are associated with criminality that is based on racist assumptions[5]; and (ii) the content of certain religious groups (e.g.

2 Differential Tweetment: Mitigating Racial Dialect Bias in Harmful Tweet Detection - https://dl.acm.org/doi/pdf/10.1145/3442188.3445875

3 Dastin, J., 'Amazon scraps secret AI recruiting tool that showed bias against women', Thomson Reuters, 10 October 2018.

4 Johnston, I., 'AI Robots Learning Racism Sexism and other Prejudices from Humans, Study Finds', The Independent, 13 April 2017.

5 AI still sucks at moderating hate speech | MIT Technology Review

Muslims) might be disproportionately blocked on the basis that they are regarded as indicative of extremism and associated with terrorism offences.

76. The problem is that the automated processes for blocking content will, most likely, operate as a "*black box*" which will make it impossible to detect such potential biases. There is no obligation under the Bill for providers to make the datasets they use to train AI public, which means there is no public or expert scrutiny of its quality or potential biases.

77. Ambit. We consider that prior restraint in the manner detailed above would fall within the ambit of Article 10 of the Convention because it amounts to a clear engagement of the right to freedom of expression.

78. Protected characteristics. Here, we understand that the potentially applicable technology has been shown to be biased against certain ethnic minority groups (race) and certain religious groups. It is possible that other protected groups may be similarly affected. For example, we are aware that AI has been shown to be biased against women in certain contexts[6].

79. Difference in treatment. The difference in treatment would be between individuals who have more of their content blocked, as compared to individuals whose content is blocked less often, if at all.

80. Differential effect. If the question was framed as indirect discrimination, there might be a disproportionate number of people in a certain protected groups that have their content censored on online platforms.

81. Objective justification. In our view, similar considerations apply here to the assessment of proportionality under Article 10 (see [60]-[71] above). In summary, it may be

[6] Dastin, J., 'Amazon scraps secret AI recruiting tool that showed bias against women', Thomson Reuters,
10 October 2018; Prates, M., Avelar, P. and Lamb, L. (2018), 'Assessing Gender Bias in Machine Translation – A Case Study with Google Translate,' CoRR abs/1809.02208, pp. 1-31.

difficult to justify the difference in treatment in light of the strong sanctions regime for non-compliance with Clause 9(2)(a) contrasted with the weak user complaints process.

82. In addition, we consider it to be particularly problematic that the decisions about what amounts to priority illegal content is: (a) a matter for the regulated providers (as non-public bodies); and (b) the AI algorithms themselves will be controlled by private parties. This means that important non-discrimination protections will not be available. For example, they will not be subject to the PSED, which requires public authorities to have due regard, inter alia, to the need to eliminate discrimination and advance equality of opportunity and for reasons set out below it is likely to be difficult, in practice, for individuals to pursue claims under the Equality Act 2010.

Claims under the Equality Act against service providers

83. While it is possible that individuals could bring claims in the county court under the Equality Act 2010 in respect of discriminatory blocking of their content, we think such claims would be difficult to mount in practice.

84. Under the EA 2010, among others, race (section 9) and religion or belief (section 10) are "*protected*". Section 29(2) prohibits discrimination by service providers. In our view, it is arguable that the content screening and blocking by regulated providers could fall within that section. Section 19(1) prohibits indirect discrimination by service providers. In principle a claim could be brought on the basis that content moderation technology operates in ways that disadvantages members of certain ethnic minority groups or religions as it disproportionately blocks content posted by or about members of such groups.

85. However, such claims are likely to be very difficult to pursue in practice. In particular:

    a. Individuals might not know (or easily be able to show) that their content has been blocked as they will not be notified of the blocking or told the reasons for it; and even if they can show material was blocked it may be very difficult to show that was the result of discrimination where the processes all occur in secret.

b. Regulated providers are likely to rely on the provision in Clause 44(2)(a) that protects them from complaints about disproportionate interferences with the right to freedom of expression if they comply with OFCOM's recommended measures (whatever those might be).
c. It is likely to be difficult for individuals to access or obtain data relating to the content moderation technology that shows that certain groups are particularly affected so as to amount an indirect discrimination claim.

86. For these reasons, we think it is likely to be difficult for individuals, in practice, to successfully vindicate their rights by bringing a claim under the EA 2010. This reinforces the need for careful scrutiny of equality considerations as the Bill is considered by Parliament.

VII. CONCLUSION

87. For the reasons detailed above, our view is that there are real and significant issues regarding the lawfulness of the current version of the Bill, and in particular the prior restraint provisions.

Dan Squires KC

Emma Foubister

Matrix

20 June 2023