Journal of Media Law

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rjml20

The Bypass Strategy: platforms, the Online Safety Act and future of online speech

Ellen Judson, Beatriz Kira & Jeffrey W. Howard

To cite this article: Ellen Judson, Beatriz Kira & Jeffrey W. Howard (25 Jul 2024): The Bypass Strategy: platforms, the Online Safety Act and future of online speech, Journal of Media Law, DOI: 10.1080/17577632.2024.2361524

To link to this article: https://doi.org/10.1080/17577632.2024.2361524

Published online: 25 Jul 2024.

JOURNAL OF MEDIA LAW
https://doi.org/10.1080/17577632.2024.2361524

Routledge
Taylor & Francis Group

The Bypass Strategy: platforms, the Online Safety Act and future of online speech

Ellen Judson[a], Beatriz Kira [b] and Jeffrey W. Howard [c]

[a]Independent Researcher, London, UK; [b]Sussex Law School, University of Sussex, Brighton, UK; [c]Department of Political Science, University College London, London, UK

ABSTRACT

In this paper, we argue that the Online Safety Act 2023 and Ofcom's guidance incentivise online platforms to adopt a 'Bypass Strategy', where they create and enforce content moderation rules that are broader than existing criminal law to bypass judgements of illegal content. This strategy aims to avoid complex legal interpretations of criminal intent and potential defences but would be unfeasible considering the volume of content on social media platforms and incompatible with automated moderation tools. We argue, however, that the Bypass Strategy, driven by the Act's focus on illegal content and by the lack of clarity in Ofcom's proposed guidance, poses a significant threat to users' freedom of expression and incentivises overremoval of legitimate speech. We offer insights that could help Ofcom to improve its guidance on how platforms should interpret such duties on moderating content and might mitigate this risk within the constraints of the Act.

ARTICLE HISTORY Received 29 February 2024; Accepted 15 May 2024

KEYWORDS Online Safety Act; social media; freedom of expression

Introduction

The debate on whether and how to subject large social media platforms to greater legal regulation has always had two dimensions, often in tension with one another.[1] On the one hand, the size and ubiquity of these online networks make it extraordinarily easy to disseminate content that causes harm. While platforms have enacted elaborate content moderation systems to restrict varieties of harmful speech, a familiar complaint (among many) is that such rules are ill-defined and in any case are enforced inadequately, exposing users (most worryingly children) to seriously objectionable content. The demand arising from this complaint, then, is that platforms

CONTACT Ellen Judson ellenejudson@proton.me

[1]See e.g. Ellen Judson, "The Online Safety Bill Position Paper" (Demos 2022) accessed 28 June 2024.

take greater steps to reduce users' exposure to harmful content (both by banning it and by making systemic changes that reduce its production, dissemination and visibility).[2]

On the other hand, the size and ubiquity of these online networks make them hugely significant fora for free expression, as their own public statements of purpose suggest[3]: these are the places where the preponderance of citizens to go share what they think, hear what others think and gather information on myriad topics. That, then, generates a countervailing worry that platforms restrict *too much* speech[4] – curtailing legitimate expression, particularly where moderation practices disproportionately affect certain groups.[5] Furthermore, such matters seem too important to be decided and overseen entirely by powerful private actors without democratic accountability.[6]

The UK's own attempt to thread the needle in accommodating these concerns was enacted, after years of debate, on 26 October 2023, when the long-awaited Online Safety Bill received Royal Assent. The Online Safety Act (OSA), stretching over 286 pages, subjects social media platforms (alongside other entities such as search engines and pornography sites) to an intricate web of new regulatory requirements, to be enforced by the UK's telecommunications regulator, Ofcom. It is fair to say that the legislation establishes substantive obligations around three principal aspects (setting aside many details, such as varyingly stringent requirements depending on company size). First, it will require platforms to take steps to combat illegal content on their networks.[7] Second, it will require platforms to take steps to protect children online from exposure to varieties of harmful content.[8] And third, it will require platforms to enforce their own terms of service consistently,[9] while allowing users various degrees of control over what kinds of content they wish to see or not see.[10]

Ofcom is already drawing up draft versions of its codes of practice offering guidance to platforms on how to live up to these duties; the

[2]See Carnegie UK, "Online Safety Act Resource Page" (Carnegie UK Trust 2023) accessed 28 June 2024.
[3]See BBC News, Zuckerberg outlines plan for 'privacy-focused' Facebook (BBC 2019) accessed 28 June 2024.
[4]See Open Rights Group, "Online Safety Bill Policy Hub" (ORG, 2024) https://www.openrightsgroup.org/campaign/online-safety-bill-campaign-hub/ accessed 28 June 2024.
[5]e.g. Christina Dinar, "The state of content moderation for the LGBTIQA+ community and the role of the EU Digital Services Act", (Heinrich-Böll-Stiftung 2021) accessed 28 June 2024.
[6]See e.g. Guy Chazan, Henry Foy and and Hannah Murphy, "Angela Merkel attacks Twitter over Trump ban" (Financial Times 2021) accessed 28 June 2024.
[7]See sections 9 and 10, OSA, on illegal content duties for user-to-user services, and sections 27 and 27, OSA, on the illegal content duties for search engines.
[8]See sections 12 and 13, OSA, on the safety duties protecting children for user-to-user services, and sections 29 and 30 on the safety duties protecting children for search engines.
[9]See, amongst others, sections 10(6), 12(10) and (11), 27(6) and 29(6) OSA.
[10]See s 15, OSA, on user empowerment duties for Category 1 platforms.

consultation conducted in early 2024 involved over a thousand pages of documents. Accordingly, there is no way to come to grips with the entirety of the legislation (and Ofcom's interpretation of it) in one discussion. Still, our aim here is nevertheless to pinpoint what we see as some central risks to citizens' rights arising from the legislation, in particular in relation to the illegal content duties. The hope is not merely to bemoan the status quo, but identify insights that could help Ofcom to improve its guidance, and – in the longer term – assist policymakers as they contemplate potential changes to the legislation.[11]

In Part I, we review the OSA's duty for platforms to identify illegal content in historical legal context. Then, in Part II, we dig deeper into the OSA's core requirements with regard to illegal speech, clarifying the difficulty of making judgements on speech's legality through at-scale automated content moderation systems. We then describe Ofcom's draft guidance on how to bypass these difficulties: enact wider, simpler content rules that restrict more speech than is legally forbidden and enforce those rules instead. In Part III, we explain the benefits of this 'bypass strategy', while calling attention to its central risk: that it endangers users' freedom of expression by incentivising the over-removal of speech. Whether this risk actually materialises in over-removal remains to be seen, but it is a risk against which Ofcom must guard. In Part IV, we speculate on how improved guidance from Ofcom might mitigate this risk within the constraints of the Act. Finally, we conclude in Part V by considering the ways in which the limitations of the current legislation might motivate future legal and policy changes.

We acknowledge that the guidance this paper is based on is still explicitly draft guidance from Ofcom, published with a view to consultation and feedback. We have focused on this guidance as it demonstrates some of the inherent conflicts that the Act presents to those tasked with enforcing it. Our aim is for this paper to not only inform immediate discussions of the details of draft guidance, but set out recommendations for a successful regulatory approach long-term.

Judging the legality of online speech: the OSA in context

Here we begin by explaining, at a relatively high level of abstraction, different regulatory frameworks' approaches requirements with regard to illegal speech on social media platforms. Along the way, we note how they compare with the UK's requirements in the Online Safety Act, both for better and for worse. We will refer to the Online Safety Act, including its

[11]Toby Helm, 'Labour Pledges to Toughen "Weakened and Gutted" Online Safety Bill' *The Guardian* (1 January 2023) <https://www.theguardian.com/technology/2023/jan/01/labour-pledges-toughen-online-safety-bill> accessed 6 May 2024.

Schedules, as well as documents from Ofcom's consultation conducted in early 2024.[12]

The UK's Online Safety Act gives the strong impression that private platforms will be required to make judgments about whether particular pieces of content is or is not illegal. The need to identify legality of content is crucial for compliance with both the illegal content risk assessment duties (Section 9) and illegal content safety duties (Section 10), both of which require companies to make judgements about the legality of content. As the Act notes:

> In making such judgements, the approach to be followed is whether a provider has reasonable grounds to infer that content is content of the kind in question (and a provider must treat content as content of the kind in question if reasonable grounds for that inference exist).[13]

The concept of social media platforms engaging in content legality assessments isn't novel. Pre-existing 'knowledge-based intermediary liability' laws already necessitate some level of legal consideration by platforms.[14] For instance, under the EU e-Commerce Directive[15] – applicable to the UK until Brexit and remaining in force under the EU Digital Services Act (DSA)[16] – platforms have enjoyed extensive exemptions from liability for illegal content, provided they lacked 'actual knowledge of illegal activity' (Article 14). While the Directive itself refrained from defining 'actual knowledge,' the Electronic Commerce (EC Directive) Regulations 2002 (SI 2002/2013) elaborated on factors courts may consider when determining such knowledge. These factors included the receipt of a valid notification detailing the unlawful nature of the activity and the specific location of the relevant information. Notably, under this 'notice-and-takedown' regime, platforms were required to assess the flagged content's legality, but (in an attempt to limit the interference with freedom of expression) were not compelled to proactively monitor all content. In fact, the Directive explicitly prohibits

[12]In particular, we will draw from Ofcom, *Protecting People from Illegal Harms Online: Volume 5 (Illegal Content Judgements Guidance)*, *Consultation* (9 November 2023) <https://www.ofcom.org.uk/__data/assets/pdf_file/0023/271148/volume-5-illegal-harms-consultation.pdf> accessed 13 May 2024 and Ofcom, *Annex 10: Online Safety Guidance on Judgement for Illegal Content*, *Consultation* (9 November 2023) <https://www.ofcom.org.uk/__data/assets/pdf_file/0025/271168/annex-10-illegal-harms-consultation.pdf> accessed 13 May 2024. Of course, in the coming months and years, this content will undoubtedly evolve in its details in response to the consultative process. But we doubt the fundamentals of Ofcom's approach will shift, and the core normative issues we discuss will remain the same.

[13]OSA, s 192.

[14]Giancarlo Frosio, 'Mapping Online Intermediary Liability' in Giancarlo Frosio (ed), Giancarlo Frosio, *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020) <https://academic.oup.com/edited-volume/34234/chapter/290264642> accessed 21 February 2023.

[15]Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') *OJ L 178, 17.7.2000, p 1–16.*

[16]Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) PE/30/2022/REV/1, *OJ L 277, 27.10.2022, p 1–102.*

Member States from imposing such a general monitoring responsibility (Article 15). The DSA, adopted by the EU in 2022, builds on the rules of the e-Commerce Directive to place on online intermediaries obligations to be more transparent about how they moderate content and to assess and mitigate a range of risks to users and institutions that their services pose. The DSA thus seeks to enhance online safety by adding a new layer of transparency and accountability obligations for online platforms, but content legality assessment is still required as the DSA requires intermediaries to act expeditiously to remove or to disable access to illegal content upon obtaining knowledge or awareness of its illegality (Article 6(1)(b)).

Another example of knowledge-based law is the controversial German Network Enforcement Act (NetzDG),[17] enacted in 2017. According to this law, social media companies must remove 'obviously illegal' content within 24 hours of a complaint, and they have up to seven days to decide on cases where the legality is not immediately apparent.[18] Unlawful content needs to be removed, with fines for a breach of this obligation reaching up to €50 million.[19] Content qualifies as unlawful under the NetzDG if it demonstrably constitutes both the *actus reus* (guilty act) and *mens rea* (guilty mind) of any of a list of 22 criminal offences outlined in the German Criminal Code (GCC). According to Wischmeyer, the benefit of relying on existing criminal statutes (rather than defining new offences for the online environment) is that platforms would already have some interpretive parameters, based on how the statutes had been applied previously applied by courts in non-digital contexts.[20]

Thus assessing content legality has long been incorporated into platform regulation legislation. Therefore, despite criticisms regarding overremoval incentives and compliance difficulties surrounding these models, platforms have over time developed some experience. YouTube's transparency report on NetzDG compliance exemplifies this:

> When we receive complaints to remove allegedly illegal content, we review each complaint carefully. If the content is in violation of local law, we will locally block the content that we identify as illegal. (…) deciding whether

[17]Network Enforcement Act (*Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken)* entered into force in 2017 to combat fake news, hate speech and misinformation online <https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html> accessed 13 May 2024.

[18]Thomas Wischmeyer, 'What is Illegal Offline is Also Illegal Online: The German Network Enforcement Act 2017' in Bilyana Petkova and Tuomas Ojanen (eds), *Fundamental Rights Protection Online* (Edward Elgar Publishing 2020); Amélie Heldt, 'Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports' (2019) 8 Internet Policy Review.

[19]Heldt (n 12).

[20]Wischmeyer also highlights the 'vague and open-ended' nature of these provisions, exemplified by the inclusion of 'insult' (s 185) despite the absence of a legal definition for such an offense. Wischmeyer (n 12). How are platforms expected to interpret whether specific content constitutes an 'insult' under the NetzDG?

content is illegal under local laws can be among the more difficult legal assessment decisions that YouTube reviewers have to make.[21]

However, the illegality assessment within the Online Safety Act differs from pre-existing requirements under knowledge-based liability laws, such as the EU e-Commerce Directive and NetzDG, in significant ways – with important implications for freedom of expression. The first key distinction lies in the OSA's nature. The OSA is grounded on the idea of duty of care and imposes duties requiring platforms to identify, mitigate and manage the risks of harm.[22] Consequently, it is not an intermediary liability law and departs from the 'notice and take-down' model employed by other legislations. Under the e-Commerce Directive and NetzDG, platforms could protect themselves from liability by removing content following an illegality assessment.

In contrast, the OSA mandates that platforms are designed to be able to identify illegal content so that they can take actions to swiftly take it down and minimise its online presence.[23] Yet the fact that individual pieces of illegal content are not taken down does not have implications for platforms' civil liability, beyond potentially demonstrating evidence that the platform's systems against illegal content are ineffective. Notably, Ofcom does require platforms to 'ensure these systems or processes are designed such that they remove illegal content swiftly where they become aware of its presence on the service'.[24] And indeed, the OSA recognises a wide range of content moderation tools beyond taking content down that could meet the risk mitigation obligation, including *ex ante* measures such as increasing friction, or using chatbot interventions.[25]

The second major difference pertains to the volume of illegality assessments required under each model. In a notice and take-down model, only content flagged as potentially breaching the law is legally required to undergo legal assessment for compliance. Conversely, the OSA's absence of a notification model for priority illegal content (section 10(3)(b)) means that the duties apply to all content on the platform, meaning that any piece of content could theoretically be subject to a legality assessment.

[21]Google Transparency Report, Removals under the Network Enforcement Law, <https://transparencyreport.google.com/netzdg/youtube?hl=en> accessed 29 February 2029.
[22]Lorna Woods, 'The Duty of Care in the Online Harms White Paper' (2019) 11 Journal of Media Law 6; Damian Tambini, 'The Differentiated Duty of Care: A Response to the Online Harms White Paper' (2019) 11 Journal of Media Law 28.
[23]OSA, s 10(3).
[24]Ofcom, *Volume 4: How to Mitigate the Risk of Illegal Harms – the Illegal Content Codes of Practice* (2023) para 12.47 <https://www.ofcom.org.uk/__data/assets/pdf_file/0022/271147/volume-4-illegal-harms-consultation.pdf> accessed 29 February 2024.
[25]Lorna Woods, 'Ofcom's Illegal Content Judgements Guidance' (*Online Safety Act Network*, 15 February 2024) <https://www.onlinesafetyact.net/analysis/ofcom-s-illegal-content-judgements-guidance/> accessed 29 February 2024.

The underspecification of OSA regulatory requirements on illegal content judgements

Social media platforms involve a firehose of speech (YouTube, for example, has 500 h of new video uploaded every minute).[26] The idea that platforms are in position to evaluate the legality of each and every piece of content posted by users is plainly beyond what is feasible. Large platforms simply cannot evaluate the legality of every post.[27]

Were the only way to discharge this task to hire an army of content moderators expertly trained in UK law, it would thus violate the 'proportionality' requirements of the OSA. Moreover, a requirement to remove all illegal speech would have the effect of incentivising platforms to err on the side of over-removal of content (e.g. removing all content for which there is a complaint of illegality, or for which moderation systems judged to be some minimal probability of illegality). Given the practical impossibility of fully accurate enforcement, this would be disastrous for free expression, and it would also be in violation of the duties OSA places on platforms to protect users' rights to freedom of expression and privacy (section 22).

Sensibly, there's an out for platforms: upon closer inspection, the duty in the OSA isn't to guarantee that *all* designated illegal content is removed. Rather, the duty is to design a system that has the outcome of reducing the presence of illegal content. To do so, services must conduct risk assessments with regard to illegal content on their service ('illegal content risk assessment duties'). Having identified the relevant risks, platforms must enact proportionate measures to reduce those risks ('illegal content safety duties') – chiefly, reducing the likelihood that users will encounter illegal content.[28]

[26]Evelyn Douek, 'Governing Online Speech: From "Posts-As-Trumps" to Proportionality and Probability' (2021) 121 Columbia Law Review 759, 791.

[27]This was also recognised when the Online Safety Bill was being discussed in the House of Lords, as expressed by Lord Parkinson of Whitley Bay 'platforms will not be penalised for making the wrong calls on pieces of illegal content. Ofcom will instead make its judgements on the systems and processes that platforms have in place when making these decisions' (HL Deb, 17 July 2023, col 2143).

[28]Section 10, OSA:

> (2) '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 …
> (3) A duty to operate a service using proportionate systems and processes designed to—
> (a) minimise the length of time for which any priority illegal content is present;
> (b) where the provider is alerted by a person to the presence of any illegal content, or becomes aware of it in any other way, swiftly take down such content.'

This is more feasible, to be sure, but it still requires judgments about illegal content overall. Even if platforms make judgments about illegality 'on a probabilistic basis',[29] they still need to operationalise criteria for judging illegality on an unprecedented scale. Ofcom recognises the challenge: 'the process of making a full assessment of whether content amounts to "illegal content" for the purposes of the Act is likely to require both more time and more legal expertise than a content moderator can reasonably be expected to have.'[30]

While acknowledging this, Ofcom's draft guidance remains ambiguous on two key points. The first is *procedural* and relates to how platforms are expected to comply with the illegal content duties, and specifically how much error platforms can afford when making assessments. How much illegal content would still be tolerable for platforms to comply with the risk mitigation obligation? The second is *substantive*, and relates to the parameters platforms should adopt to assess the legality of content. How are platforms to determine whether there are reasonable grounds to infer illegality? As we will show, the answers provided by Ofcom to each of these questions seem to be in conflict.

On the procedural question, it is possible to interpret Ofcom's draft guidance in two conflicting ways. The first interpretation is that getting the decision about the legality of content right is very important, since a substantive part of the guidance is focused on assessing the legality of specific pieces of content.[31] The second interpretation is that judgements about specific pieces of content are not that important, since Ofcom clarifies that compliance with illegal content duties will not be assessed based on the presence of individual pieces of content, but rather on the systematic efforts undertaken by companies to mitigate risks.[32]

Obviously, we cannot gauge the reasonableness of the expectations placed on platforms without knowing exactly what those expectations are. Consider this passage: 'Service providers must ensure these systems or processes are designed such that they remove illegal content swiftly where they become

[29] Ofcom, *Annex 10: Online Safety Guidance on Judgement for Illegal Content* (n 6) para A1.15.
[30] ibid A1.17.
[31] According to volume 5, 'the Act requires us to provide guidance to services about how they can judge whether a piece of content is likely to be illegal', Ofcom, *Volume 5: How to Judge Whether Content is Illegal or Not? (Illegal Content Judgements Guidance)* (n 6) p. 4. <https://www.onlinesafetyact.net/analysis/ofcom-s-illegal-content-judgements-guidance/> accessed 29 February 2024. Further, 'when a service is making an illegal content judgement each piece of content will need to be considered on a case-by-case basis with reference to the state of the mind requirements of the offence and any available defences as prescribed by' Ofcom, *Annex 10: Online Safety Guidance on Judgement for Illegal Content* (n 6) para A1.71.
[32] 'When services conduct risk assessments and implement measures to comply with their safety and other duties, they are likely to be dealing with content in bulk, as opposed to making an assessment on an individual piece of content. Services should anticipate that some of the content they hold is likely to be illegal content, but can do this on a probabilistic basis', Ofcom, *Annex 10: Online Safety Guidance on Judgement for Illegal Content* (n 6) para A1.15.

aware of its presence on the service'.[33] This makes it clear the requirement is not absolute (not every piece of illegal content must be removed), but it does not make it clear what the threshold of success is at which a system is deemed effective enough, nor indeed how that threshold can be measured by the regulator (who also cannot be conducting mass illegal content assessments!). The indication from recommendation 4C (Annex 7) is that the provider themselves should set 'performance targets' for its content moderation function, covering at least: (a) the time that illegal content remains on the service before it is taken down; and (b) the accuracy of decision making. The draft guidance also states that in setting such targets platforms should 'balance the desirability of taking illegal content down swiftly against the desirability of making accurate moderation decisions.'[34] Does this mean that platforms can adopt whatever performance standards they like, no matter what they are? Are any ways of striking the desired balance unreasonable, and if so, how will platforms know? These questions do not have clear answers. In contrast, the answer to the second, substantive question, related to parameters to assess illegality, is provided by the Act. Still, it is difficult to reconcile this answer with the systemic view developed in Ofcom's guidance. The OSA requires platforms to have 'reasonable grounds to infer' that content is illegal. It notes:

> Reasonable grounds for that inference exist in relation to content and an offence if, following the approach in subsection (2), a provider – (a) has reasonable grounds to infer that all elements necessary for the commission of the offence, including mental elements, are present or satisfied, and (b) does not have reasonable grounds to infer that a defence to the offence may be successfully relied upon.[35]

The challenge becomes even more daunting once we consider the fact, obvious to criminal lawyers, that crimes typically necessitate not just some criminal action (the *actus reus*) but also some mental element – such as intention or foresight – on the part of the agent (the *mens rea*). Yet platforms' content moderation systems are notoriously ineffective at inferring the mental states of users. This is, in part, because of the limitations of the automated tools that platforms use. But even human moderators lack the requisite context to make confident judgments about users' mental states in the limited time they would have to make them;[36] after all, they are asked

[33]Ofcom, *Volume 4: How to Mitigate the Risk of Illegal Harms – the Illegal Content Codes of Practice* (n 18) para 12.47.

[34]Ofcom, *Annex 7: Illegal Content Codes of Practice for User-to-User Services* (2023) para A4.11–A4.12 <https://www.ofcom.org.uk/__data/assets/pdf_file/0022/271165/annex-7-illegal-harms-consultation.pdf> accessed 29 February 2024.

[35]OSA, s 192(6).

[36]For a detailed discussion of the work done by human content moderators, see Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 Harvard Law Review 1598.

quickly to judge snippets of text, not rich testimony about speakers and their conduct.[37]

Consider the case of the 'Robin Hood Airport' tweet, in which a man joked about blowing up the airport in January 2010. He was initially convicted of a communications offence, and after a series of appeals won his case in 2012, over two years later.[38] If the justice system changes its mind about the legality of one tweet, and takes 2 years to do so, it is hardly reasonable to expect social media platforms to effectively decide so in a matter of seconds. For example, in sharing a personal story of attempted suicide, is the speaker permissibly raising awareness about his own experiences, or illegally encouraging self-harm?

Thus the apparent requirement that platforms make judgments about what content is illegal is enormously fraught. Even if platforms are not required to make judgements about *all* content, it still needs a system designed to sort legal from illegal content – where the latter is understood as content for which there is a reasonable inference of illegality. But what, exactly, counts as a reasonable inference that some post is illegal? On this point, the OSA punts the issue to Ofcom, whom it tasks with the job of providing guidance to the platforms (section 193, OSA). We already have the first indications of its views on this question. Ofcom recognises fully that judging *mens rea* online will be particularly difficult, but the regulator also realises that it cannot put this aside because of the Act's requirement that the mental element be considered as part of the 'reasonable grounds to infer' assessment.[39] Yet, given the nature of this assessment, certainty (i.e. 'beyond reasonable doubt') is not what is required: 'Reasonable grounds to infer is not a criminal threshold, and there are no criminal implications for the user if their content is judged to be illegal content against this threshold.'[40] Ofcom continues:

> We recognise that in some cases, particularly where there may be many reasons for a person to do as they have done, it will likely never be possible

[37] Seeming to recognise this point, the OSA notes that 'judgements are to be made on the basis of all relevant information that is reasonably available to a provider' (OSA, s 192(2)).

[38] 'Robin Hood Airport Tweet Bomb Joke Man Wins Case' *BBC News* (27 July 2012) <https://www.bbc.co.uk/news/uk-england-19009344> accessed 29 February 2024.

[39] '26.43 We acknowledge that inferences about state of mind are particularly difficult in online situations, where contextual clues are often not apparent and, for example, what would be an obvious joke or piece of sarcasm in an offline context might not appear so obvious when online. We also acknowledge that conclusions about state of mind in criminal cases are nuanced, and usually draw upon an extensive suite of evidence which is not reasonably available to a service moderating a single piece of content. However, neither Ofcom nor in scope services can put aside the state of mind or "mental element" requirement as this is a part of the "reasonable grounds to infer" threshold, as established by the Act' Ofcom, *Volume 5: How to Judge Whether Content is Illegal or Not? (Illegal Content Judgements Guidance)* (n 6) para 26.43; See also Ofcom, *Annex 10: Online Safety Guidance on Judgement for Illegal Content* (n 6) para A1.43.

[40] Ofcom, *Annex 10: Online Safety Guidance on Judgement for Illegal Content* (n 6) para A1.46.

> to reach firm conclusions about a poster's state of mind. However, the Act does not require proof to the criminal standard, and therefore neither will Ofcom when assessing a service's compliance.[41]

It's worth pausing to reflect on why the use of a civil standard is apt. The rationale for a balance-of-probabilities standard in civil law reflects the fact that the stakes in civil law tend to be lower.[42] In this context, what is at stake (beyond fines to companies) is whether users' posts will remain up, not whether they will be sent to prison.[43] This isn't to diminish the stakes: removal of posts (and downstream consequences such as account suspension) involve important communicative interests. But the stakes are nevertheless lower than is typical in the criminal context. Moreover, given the difficulty in evaluating the legality of online speech, insisting upon the criminal 'beyond reasonable doubt' standard would have the effect of dramatically lowering how much harmful speech is removed – thereby failing to strike the right balance between respecting speech and preventing harm. In addition, there seems to be a different standard when it comes to the assessment of illegal *conduct* or behaviour, which is not based on the content itself, but on the contextual information available.[44] Ofcom states that 'inferences may reasonably be made from the contextual information that is available to a moderator on a case-by-case basis' and that the 'the conduct or behaviour criteria of an offence may be inferred to be present and satisfied based on the likelihood that this is the case'.[45] Still, while the use of a civil standard is defensible, it alone does not obviate the gargantuan nature of the task at hand.

The workaround: the bypass strategy

These factors mean that platforms have a seemingly impossible task: implement a highly accurate, legally nuanced system of content moderation which works at scale.

[41] ibid p 15.

[42] It is worth noting that we are assuming a civil balance-of-probabilities (i.e., >50%) standard here. And yet in the Ofcom guidance, Ofcom notes that '"reasonable grounds to infer" is a new legal threshold' Ofcom, *Volume 5: How to Judge Whether Content is Illegal or Not? (Illegal Content Judgements Guidance)* (n 6) para 26.4. But that is confusing, since clearly the civil law standard is not a 'new' threshold – so Ofcom should spell this out further.

[43] Ofcom stresses that the OSA does *not* generally compel platforms to report to police users whose speech is removed on grounds of illegality (with some exceptions – e.g., CSAM content).

[44] According to Lorna Woods, 'The illegal content safety duties are triggered by content linked to a criminal offence, not by a requirement that a criminal offence has taken place. Indeed, the Consultation states that it is not the purpose of the regime to make decisions on whether a criminal offence has taken place'. Lorna Woods, 'Ofcom's Illegal Content Judgements Guidance' (Online Safety Act Network, 15 February 2024) <https://www.onlinesafetyact.net/analysis/ofcom-s-illegal-content-judgements-guidance/> accessed 29 February 2024.

[45] Ofcom, *Volume 5: How to Judge Whether Content is Illegal or Not? (Illegal Content Judgements Guidance)* (n 6) para 26.82.

In response, Ofcom offers an intriguing and quite fundamental workaround: it strikingly interprets the OSA *not to require that platforms make judgments of illegal content at all*! Despite the prodigious energy put into elaborating how platforms should judge whether content is illegal (Annex 10, which covers this point, is 390 pages), platforms don't actually need to do it after all. As Ofcom explains:

> there is nothing in the Act that requires services to make illegal content judgments, so long as the application of that service's own terms and conditions is sufficient to secure compliance with the duties in the Act in other ways. For example, if the service's own terms and conditions of use prohibit content that is wider than the definition of illegal content under the Act, then the service provider would be considered to have fulfilled its legal duties regarding takedown so long as it applied these terms and conditions properly.[46]

So, for example, consider sexual content. Only some sexual content is illegal, and thus is the kind of content implicating services' illegal content duties. Yet platforms can bypass the thorny question of distinguishing *illegal* sexual content from *legal* sexual content, by banning (as some already do) *all* sexual content.[47] And indeed this is precisely the strategy that Ofcom anticipates services to follow:

> It is our assumption that most services will take the approach explained above as it allows them to freely moderate content based on their own terms of service (or equivalent), rather than having to make illegal content judgements based on our guidance.[48]

Ofcom suggests that the best way through is to just define 'wider' rules that drop any attention to *mens rea* or defences, sparing oneself that thorny exercise. In other words, platforms could adopt something like a strict liability account of their rules, whereby no attention to speakers' mental states (or available defences) is considered whatsoever. Ofcom more or less suggests precisely this in its consultation materials.[49]

This is permitted under the Online Safety regime on the assumption that the outcome is still successful reduction of illegal content. Call this strategy – to bypass judgements of illegal content – the *Bypass Strategy*.

Ofcom are provisionally recommending this strategy as the alternative to making illegal content judgements:

[46]Ofcom, *Annex 10: Online Safety Guidance on Judgement for Illegal Content* (n 6) para A1.3.
[47]ibid A1.18.
[48]Ofcom, *Volume 5: How to Judge Whether Content is Illegal or Not? (Illegal Content Judgements Guidance)* (n 6) para 26.19.
[49]Strikingly, Ofcom even concludes that, despite Parliament's explicit insistence that defences be considered, defences should play no practical role in platform decisions. 'We believe that general defences are unlikely to be relevant to a service's illegal content judgments as it is difficult to imagine circumstances in which services would have reasonable grounds to infer that they arise. As such, we propose not to outline these general defences in the guidance' ibid 26.84.

> We are provisionally recommending that all regulated U2U services should have systems or processes designed to take down illegal content of which they are aware swiftly.
>
> For this purpose, when a service has reason to suspect that content may be illegal content, it should either:
>
> 35 a) make an illegal content judgement in relation to the content and, if it determines that the content is illegal content, take the content down swiftly; or
>
> b) where the provider is satisfied that its terms and conditions for the service prohibit the types of illegal content defined in the Act which it has reason to suspect exist, consider whether the content is in breach of those terms of service and, if it is, take the content down swiftly.[50]

What should we make of the Bypass Strategy? Here we catalogue the benefits of enabling platforms to bypass illegal content judgement, before noting some risks of this approach.

The first benefit is a benefit for companies themselves. As we have noted, it is enormously fraught for platforms to make illegal content judgements. Indeed the strategy seems to constitute an implicit admission from Ofcom that tasking platforms with making judgments about illegal content was always a fraught matter. The elaborate guidance offered by Ofcom (Annex 10) on this point underscores this point, given the difficulties of judging *mens rea*. The difficulty, we think, is quite fundamental, and it concerns the completely different nature of the criminal justice system, on the one hand, and online content moderation, on the other: The definitions of crimes under the law are not specified with an eye toward their administrability by a mass system of largely mechanised enforcement;[51] thus the prospects of designing a content moderation system that will tightly track the contours of UK criminal law (or any other jurisdiction) are slim. That slimness, in turn, generates legal exposure for platforms, who are bound to fail in such an effort and potentially incur hefty fines as a result.

The second benefit is that such an approach enables companies to double down on enforcing their global content rules effectively, rather than operating a two-tiered system whereby they enforce their own rules *plus* enforce a jurisdiction-specific set of rules emanating from the domestic criminal law of a particular polity. For example, the EU Digital Services Act's takedown requirements are indexed to illegal content as defined in other EU laws or

[50]Ofcom, *Volume 4: How to Mitigate the Risk of Illegal Harms – the Illegal Content Codes of Practice* (n 18) para 12.18.

[51]On the adoption of automated content moderation by platforms, see Hannah Bloch-Wehba, 'Automation in Moderation' (2020) 53 Cornell International Law Journal 41; Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 205395171989794.

domestic law of the various EU member states, producing precisely such two-tiered burden for companies.[52] Ofcom is clearly keen to avoid that.

Especially if platforms commit themselves to a global system of rules under the norms of international human rights law (as Meta and its Oversight Board have committed to)[53], it is needlessly complicated (perhaps) to muddle that commitment with a discrete programme to become (in a way) part of the domestic law enforcement apparatus of every state in which a platform operates. More tentatively, this may make it easier to answer the objection that the UK is inadvertently setting a bad precedent, encouraging authoritarian states with draconian speech laws to enlist platforms in those censorial laws' enforcement.

But the Bypass Strategy also comes with serious risks, given the clear incentive it creates. On the one hand, it may enable platforms to do nothing new at all. Most major social media platforms ban certain kinds of illegal content explicitly in their terms of service, and thereby already have a system in place to tackle these harms arising. A platform minded to do so could plausibly claim to Ofcom that it need not change anything about its policies, despite the Online Safety Act being brought in specifically to address the failures of platforms to self-regulate in these ways effectively. Identifying a systemic failure of compliance, therefore, would become Ofcom's burden to bear, which would mean identifying illegal content occurring at scale on the platform: a task even harder for an under-resourced regulator than a platform.[54] It also does not do any good for the regulation to become a known fiction – one in which platforms, Ofcom and the public know they are not enforcing the letter of the Act, because the demands of the Act are too strenuous to either uphold, identify or enforce against. Without a more clearly articulated measure of success from Ofcom, it remains very difficult to gauge whether a platform has lived up to its obligation. We saw above that this was true for platforms pursuing the normal route of identifying illegal content as such; it remains the case even if pursuing the Bypass Strategy.

On the other hand, for platforms wanting to take a more cautious approach, Ofcom's guidance for platforms arising from the strategy is crystal clear: *adopt platform rules that restrict more speech than is illegal, and you can bypass judgments about what exact content counts as illegal.*[55]

[52]Article 3(h), DSA, '"illegal content" means any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law'.

[53]Miranda Sissons, "Our commitment to human rights" (Meta 2021) <https://about.fb.com/news/2021/03/our-commitment-to-human-rights/> accessed 28 June 2024.

[54]On Ofcom's capacity constraints, see Lisa-Maria Neudert, 'Regulatory Capacity Capture: The United Kingdom's Online Safety Regime' (2023) 12 Internet Policy Review.

[55]It may seem that platforms must still make legal judgments about individual pieces of content in order to comply with the takedown duty – i.e., the duty to remove illegal content reported to it. But

Given how fraught such judgments are (as we have already indicated), this incentive will be a powerful one. Even platforms that prefer to hew closely to the criminal law will have a clear incentive to err on the side of caution and adopt expansive terms of service.

Now, incentivising platforms to remove more speech than what is already illegal is not *necessarily* objectionable. After all, one of the insights of earlier versions of the Online Safety Bill was a recognition that some content that is properly legal offline may become harmful when aggregated and amplified online – such that platforms should be responsible for managing the risks of such (otherwise legal) harmful content. Adopting rules wider than UK criminal law are not, therefore, *automatically* cause for concern from a free speech perspective. Platforms already remove more speech than is currently illegal, and indeed this is often necessary to maintain not only safety but usability of a platform (by weeding out spam, for instance).[56]

The real danger, in our view, is that platforms will expand their terms of service (or their interpretation thereof) beyond what they judge necessary or desirable for their users and service, in order to reduce any legal risk that they will be charged with flouting their illegal content duties. Given the difficulty in identifying what the Act actually requires platforms to do, highly risk-averse platforms may opt for this approach – ratcheting up their content moderation systems accordingly. While it is difficult to predict whether this will occur, it would be a serious case of unintended consequences: after all, amendments to the then Bill focused it on illegal content duties precisely to avoid the risk of platforms being incentivised to over-moderate legal speech.[57]

If it turned out that platforms responded to the legislation by deliberately over-removing legitimate speech, it would raise the crucial question of whether platforms are respecting their legal duties *vis a vis* users' freedom of expression. After all, the OSA instructs services as follows: 'when deciding on, and implementing, safety measures and policies, a duty to have particular regard to the importance of protecting users' right to freedom of expression within the law' (section 22(2) OSA). In theory, this ought to serve as a counterbalancing measure to the threat of platforms' over-removal of speech which would have been both legal and permitted but for compliance with the illegal content duties. In a striking note, however, Ofcom complicates the picture:

according to Ofcom, this is not so: it suffices to enforce one's existing terms of service, just in case one's rules are more expansive than what counts as illegal content.

[56]See Alex Krasodomski-Jones, 'Everything in Moderation: Platforms, Communities and Users in a Healthy Online Environment' (Demos 2020) <https://demos.co.uk/research/everything-in-moderation-platforms-communities-and-users-in-a-healthy-online-environment/> accessed 29 February 2024.

[57]On this point, see John Woodhouse, Lorraine Conway and Sally Lipscombe, *Research Briefing: Online Safety Bill: Progress of the Bill* (House of Commons Library 2023). It is striking that Ofcom's discussion of freedom of expression in Volume 5 is restricted to a quarter of a page Ofcom, *Volume 5: How to Judge Whether Content is Illegal or Not? (Illegal Content Judgements Guidance)* (n 6) para 26.126–26.129.

> Potential interference with users' freedom of expression arises where content is taken down because the service considers it to be illegal content, particularly if that judgement is incorrect. As set out above, however, our starting point is that Parliament has determined that services should take proportionate steps to protect UK users from illegal content. Of course there is some risk of error in them doing this, but that risk is inherent in the scheme of the Act.[58] (…) A greater interference would arise if the service, because of the Act, chose to adopt terms of service which defined the content it prohibited more widely than is necessary to comply with the Act. However, it remains open to services as a commercial matter (and in the exercise of their own right to freedom of expression), to prohibit content that is not or might not be illegal content, so long as they abide by the Act. Nothing in this option asks that services take steps against any content other than illegal content. Services have incentives to meet their users' expectations in this regard, too.[59]

As such, there appears to be no freedom of expression concern from the regulator about platforms adopting the Bypass Strategy. This is highly unfortunate. Even if platforms are at liberty to moderate more speech than is illegal,[60] it doesn't follow that the state is permitted to strongly incentivise them to do so. When Ofcom states that '[n]othing in this option [i.e. the Bypass Strategy] asks that services take steps against any content other than illegal content,'[61] it is verging on implausible. The Herculean feat of reliably catching illegal content, and distinguishing it from legal content, is sufficiently great that platforms face powerful legal incentives to adopt the Bypass Strategy.[62] To put it bluntly, threats of state coercion are foreseeably incentivising the most important forums of public discourse to shut down legitimate expression. This is plainly an issue of free expression.[63]

How might this danger play out in practice? What legitimate speech, in other words, might wind up on the chopping block? Recall the fact that, were the platforms to attempt only to take down illegal content, they would need to implement an at-scale system for judging whether the relevant *mens rea* condition is satisfied, and whether legal defences are unavailable,

[58] Ofcom, *Volume 4: How to Mitigate the Risk of Illegal Harms – the Illegal Content Codes of Practice* (n 18) para 12.64.

[59] ibid 12.67.

[60] Some will dispute this, given their central role for public discourse, which makes them closer to state actors. We do not rely on any such controversial claim here. Our point is that, while it is fine for private intermediaries to take down legal content, *they must not be doing it because they were incentivised to do so by the state.*

[61] Ofcom, *Volume 4: How to Mitigate the Risk of Illegal Harms – the Illegal Content Codes of Practice* (n 18) para 12.67.

[62] With thanks to Reviewer 2: it is worth noting that Ofcom's guidance requires platforms to monitor their success at quickly and accurately removing illegal content. See Ofcom, *Annex 7: Illegal Content Codes of Practice for User-to-User Services* (n 28) recommendation 4C, meaning that even if the Bypass Strategy is used in order to make content moderation decisions, at some point platforms will still need a system to make rough illegal content judgements in order to facilitate this monitoring.

[63] In the American context, the Supreme Court has long held that state action incentivising intermediaries to suppress legal, protected speech – as an unintended side-effect of requiring them not to host illegal, unprotected speech–violates the First Amendment. *Smith v California* 361 U.S. 147 (1959).

when evaluating posts. Even with a reduced epistemic standard from beyond-reasonable-doubt, such a task is Herculean, and guaranteed to be riddled with error. Ofcom's guidance highlights that particularly in relation to offences of hate, abuse or harassment, context is particularly relevant and freedom of expression particularly at risk of being accidentally infringed upon.[64]

So, imagine a platform enacted a policy on threatening language whereby any speech that takes the form of a threat will be removed, regardless of the speakers' mental states (e.g. even if the threat is plainly sarcastic). Or imagine a platform enacted a policy whereby *any* discussion about self-harm is removed – regardless of whether the speaker was encouraging self-harm or raising awareness. But such a move would be disastrous for free expression. The reason the law cares about *mens rea* and defences for speech crimes is precisely that it would plainly violate freedom of expression to punish speakers in cases where they didn't satisfy the relevant *mens rea*, or in which they enjoyed a relevant defence. *Thus the Bypass Strategy incentivises platforms to do what the state itself could never do*.

The pressure to adopt wider terms of service goes beyond the incentive to strip out any concern for *mens rea* or defences. Simply for convenience and accuracy, consider the fact that assisting illegal immigration is a priority offence under the Act (Schedule 7, para. 23, OSA). This caused controversy when it was announced that content which portrayed Channel crossings in a 'positive' light'[65] could come under this offence.

If platforms' terms and conditions stated that any content which assisted illegal immigration would be removed, under the provisions of the Act, they would have to have systems in place designed to remove all and *only* (section 71(1), OSA) content which did in fact assist illegal immigration – with potential consequences if they either fail to remove the prohibited content, or fail to leave up the permitted content. If, however, a platform puts in their terms of service that 'no-one may discuss anything about illegal immigration, such as Channel crossings', and automatically remove content mentioning certain keywords, they would be likely to successfully remove a lot of the illegal content, and also maintain their terms and conditions and avoid sanction. Much effort has gone into platforms improving their terms and conditions over the last few years to take into account context and nuance: this regulation risks incentivising the reversal of this, for fear of increasing their own moderation error rates by having sensitive and detailed moderation rules. This would then mean that the state attempt to censor illegal

[64]See, for example, Ofcom, *Annex 10: Online Safety Guidance on Judgement for Illegal Content* (n 6) para A1.61.

[65]Diane Taylor, 'A Ban on "Positive" Videos Won't Stop the Channel Crossings, but It May Well Cause More Tragedies' *The Guardian* (19 January 2023) <https://www.theguardian.com/commentisfree/2023/jan/19/ministers-ban-videos-channel-crossings-small-boats> accessed 29 February 2024.

content about immigration would have had the knock-on effect of censoring content about refugees, asylum seekers and immigration in general.[66]

Not being able to discuss immigration might not seem inherently a concern for freedom of expression: for instance, a community forum dedicated to a particular topic might reasonably ban discussion of another topic[67] (such as r/politics, which prohibits submissions about non-US politics[68]). But if the pattern repeats across all topics which could somehow be linked to conversation which might constitute an offence, and across the most major platforms which provide forums for public discussion[69] – immigration, weapons, suicide, prostitution and anything mentioning a protected characteristic for which there are hate speech laws – the consequences for citizens would be that their ability to express themselves in the digital public square be vastly curtailed.

In raising these concerns, we must once again stress: none of this may actually happen! After all, platforms face countervailing public pressure to leave up speech. And it's important to recognise that platforms have enacted rules that *do* try to take into consideration the intentions of speakers (e.g. on Meta, posting images of human rights abuses to raise awareness can be permissible whereas posting them to promote thems is not).[70] The problem is that existing efforts to incorporate speakers' mental states are enormously fraught, and are liable resulting in quite high error rates. The issue is that, once platforms face *legal* pressure to reduce such error rates, they will simply flatten the requirements – forfeiting imperfect but aspirationally nuanced enforcement in order to reduce legal exposure. It is that problem against which Ofcom must guard.

We must also recognise that Ofcom is limited in what it can do to mitigate this concern; after all, its hands are largely tied by the wording of the Act itself. It cannot operate outside the parameters set up by the Act – as such, any proposals to reform the essential nature of the duties around illegal content, while potentially more satisfying, cannot be brought in by Ofcom themselves, they would require Parliament to act. We thus present three steps that Ofcom should take to mitigate the risks to freedom of expression while maintaining the essential force of the illegal content duties.

[66]Monica Horten, 'Could Debate on Immigration Be Suppressed?' (*Open Rights Group*, 1 September 2022) <https://www.openrightsgroup.org/blog/could-public-debate-on-immigration-be-suppressed-by-the-online-safety-bill/> accessed 29 February 2024.

[67]See Alex Krasodomski-Jones, 'Everything in Moderation: Platforms, Communities and Users in a Healthy Online Environment' (Demos 2020) <https://demos.co.uk/research/everything-in-moderation-platforms-communities-and-users-in-a-healthy-online-environment/> accessed 29 February 2024.

[68]<https://www.reddit.com/r/politics/about/> accessed 29 February 2024.

[69]See Jack M Balkin, 'How to Regulate (and Not Regulate) Social Media' (2021) 1 Journal of Free Speech Law 71.

[70]Meta, "Violent and graphic content" (Transparency Center 2024) <https://transparency.meta.com/en-gb/policies/community-standards/violent-graphic-content/> accessed 27 June 2024.

Mitigating the risks of the bypass strategy

In order to reduce the risk to freedom of expression identified, Ofcom should change the illegal contents judgement guidance to be something a platform can actually do; and set the error bar low.

First, we think Ofcom should reconceive the Illegal Contents Judgement Guidance as currently drafted. The perverse incentives arise because platforms face sanctions if they either fail to correctly enforce their terms of service, or fail to correctly apply the illegal contents judgement guidance. As we saw, one way to avoid both of these failures is to make terms of service wider and simpler, to remove context or nuance dependencies which increase the chance of platform judgement error in its removal. If, however, it was less difficult to meet the requirements of the illegal content judgement guidance, platforms would be less incentivised to avoid it. The way to accomplish this, we suggest, is to offer more precise guidance on how much (and what type of) illegal content it is tolerable for platforms to have and how this will be practically assessed by the regulator. For political reasons, legislators are inclined to say 'none!' and avoid actually confronting the trade-off here. But Ofcom's independence puts it in a strong position to take a clear stand on this issue.

Second, Ofcom should set out specific guidance alongside the illegal contents guidance on how platforms should safeguard freedom of expression if undertaking the Bypass Strategy. The indication of the guidance is that there is no possible level at which platforms' terms of service would be taken to breach users' freedom of expression (bar, for instance, if platform terms of service breached other obligations, such as disproportionately impacting the expression of a certain group). Ofcom's recommendations for safeguards on content moderation processes include: recommending setting performance targets to ensure that content moderation processes are accurate; recommending providing adequate training to moderators; and having appeals processes for complaints.[71] These do not mitigate the risks of the Bypass Strategy, as it only seeks to reduce inaccurate content moderation, rather than accurate content moderation under a wide content moderation strategy. We suggest that Ofcom should produce additional guidance on safeguarding freedom of expression under the Bypass Strategy: and acknowledge this threat explicitly, rather than (as currently) asserting that there is no freedom of expression concern from platforms' own commercial decisions. Indeed, given the incentives that platforms face to adopt the Bypass Strategy, we think Ofcom should require Category 1 platforms to address how they are mitigating risks to freedom of expression.

Finally, Ofcom should produce a non-binding draft terms of service which it considers to be sufficient to discharge a platforms' illegal content

[71] Ofcom, *Annex 7: Illegal Content Codes of Practice for User-to-User Services* (n 28) para A4.5.

duties without going above and beyond them. This would give platforms a clear standard to follow if they do choose the Bypass Strategy and avoid the need for them to define their terms of service excessively widely in order to encompass any ambiguous areas. In offering these suggestions, we emphatically are not rejecting the Bypass Strategy. Ofcom is entirely correct that the alternative – requiring platforms to make judgements about the legality of content – raises insurmountable hurdles. Compared to that option, which is wholly infeasible to administer, the Bypass Strategy is superior. Ofcom is reasonable, we surmise, to encourage platforms to draw wider rules than what the criminal law tightly prohibits. But it doesn't follow that any rules will suffice. Under the current guidance, Ofcom is essentially suggesting to platforms: 'Draw whatever rule you want, so long as it encompasses presently illegal speech.' This approach is certain to sweep up huge swaths of legitimate speech, for the sake of ensuring that illegal content is caught. It is this that needs to be guarded against. The current Act offers Ofcom some resources for incentivising platforms to guard against the excessive over-removal of legitimate speech (albeit not enough). The steps we have sketched in this section offer a partial, immediate path.

However, even if Ofcom were to radically change their approach from their current guidance, avoiding some of these inherent tensions is difficult while remaining within the letter of the remit of the OSA itself. This should be a lesson to policymakers that regulating too much for what would be good in principle rather than in practice, means having to either retroactively change legislation, or grant regulators more freedom of interpretation than Parliament may be willing to make a habit of.[72]

Conclusion: the wider situation

This paper has focused on one particular corner of online speech regulation in the UK – the duty to combat illegal speech under the Online Safety Act regime – unpacking its difficulties and risks. But there are broader lessons, we think, to be learned about the governance of online speech flowing from this particular case.

The first lesson is that it is extraordinarily difficult to apply rules tailor-made for the habitat of judge-and-jury-staffed criminal trials, to the incredibly different world of governing online speech. Any attempt at the kind of nuance required in the world of criminal law (factoring in speakers' mental states and available defences) will result in much higher rates of error when

[72]See, for instance, the recommendation for an ongoing Joint Committee which could 'provide a greater level of democratic accountability for Ofcom' to guard against regulatory overreach. See Joint Committee on the Draft Online Safety Bill, *Draft Online Safety Bill* (House of Lords and House of Commons, 14 December 2021), section 433, p 124, <https://committees.parliament.uk/publications/8206/documents/84092/default/> accessed 7 May 2024.

such judgments (or analogous ones within platforms' rules) are made by human and machine moderators. But if platforms are exposed to legal liability for such errors, they will be incentivised to broaden and simplify their rules in ways that are deleterious for free speech.

The second lesson is that a focus on illegal speech can cause us to miss the forest for the trees – focusing on individual units of content, rather than the broader system. The proposals underpinning the OSA grew out of the thought that platforms have a duty of care to design their spaces in ways propitious for constructive free expression and the safety of users – not by playing whack-a-mole with individual content but focusing on the broader ecosystem. Ofcom has interpreted the act in ways that emphasise content takedown as the most significant policy lever. But there is a broader interpretation that considers the whole ecosystem – in particular, the ways that platform architecture might incentivise or facilitate the production and distribution of prohibited content. It is, we think, possible for Ofcom to take this wider view, and it should.

That leads to a third and final lesson. As noted above, some content that is properly legal becomes harmful (and so, we think, unprotected speech) when aggregated and amplified online. It can be reasonable to hold platforms accountable for reducing the risks posed by such speech, given that its aggregation and amplification is the result of platforms' own architectures. That was the insight underlying the unfortunately named 'adult safety duties' that appeared in earlier versions of the then Online Safety Bill, and is the approach adopted in the EU Digital Services Act, which requires very large platforms and search engines to assess and mitigate risks beyond illegal content – including negative effects to fundamental rights and to civic discourse and electoral processes.[73] The general point is that some content poses accumulated harms that platforms should take responsibility for mitigating; yet it would be misguided to criminalise each individual post. That insight remains valid. Future policymakers would do well to take this point to heart as they revisit digital policy.

Acknowledgements

We are grateful to anonymous referees for helpful feedback on this piece, as well as to Hedvig Schmidt for her comments as editor of this special issue. We are also thankful to UKRI for research funding, which enabled Beatriz Kira and Jeffrey Howard to work on this article (UKRI grant MR/V025600/1).

Disclosure statement

No potential conflict of interest was reported by the author(s).

[73] Art 34 (1)(b) and (c), DSA.

Notes on contributors

Ellen Judson is a digital policy and disinformation specialist, who has worked extensively in NGOs on online harms and digital rights research and policy.

Beatriz Kira is a Lecturer in Law at the University of Sussex and a Research Fellow at UCL's Digital Speech Lab. She is also a Visiting Scholar at the Blavatnik School of Government, University of Oxford.

Jeffrey W. Howard is Associate Professor of Political Philosophy and Public Policy at University College London, where he is Director of the Digital Speech Lab. He is also Senior Research Associate at the Institute for Ethics in AI at Oxford University.

ORCID

Beatriz Kira http://orcid.org/0000-0002-7078-8193
Jeffrey W. Howard http://orcid.org/0000-0002-6521-9228