Analysis: Ofcom's Illegal Content Judgements Guidance

Authored by Professor Lorna Woods OBE, University of Essex
Contact at **hello@onlinesafety.net**

Issue

There are a number of concerns in relation to the definition of illegal content and the Illegal Content Judgements Guidance (Annex 10) proposed by Ofcom in its consultation on the Online Safety Act Illegal Harms duties which in effect define the scope of the regime relating to illegal content:

- How Ofcom's approach fits with "safety by design" principles;
- The degree to which the Guidance is focussing on identification of criminal conduct rather than content associated with a criminal offence;
- Standards of proof should be civil not criminal – the regime is regulatory;
- Impact on protection of human rights.

Of course, Ofcom is constrained by the provisions of the Act, but as we set out below, those provisions are open to a number of interpretations. Ofcom's chosen interpretation and the proposals that flow from it in the draft Guidance will lead to a restrictive focus on takedown of content rather than a more systemic approach.

What the Act says

Section 59 defines "illegal content", specifying content the use of which "amounts to" or the possession, viewing, publication or dissemination of which "constitutes" a relevant offence. A relevant offence is either a priority offence (listed in Schedules 5, 6 and 7) or one that satisfies the criteria in s 59(5).

Given that content on its own is not a relevant criminal offence, but rather certain behaviours (identified in s 59(3)(a)-(c)) linked to the content, and also that there are difficulties in identifying the other elements of the offence, section 192 sets out "Providers' judgements about the status of content" and Section 193 requires Ofcom to produce Guidance on it.

Section 192 (1) identifies the purpose of the section, which is to set out the approach to be taken where:

> (a) a system or process operated or used by a provider of a Part 3 service for the purpose of compliance with relevant requirements,
>
> (b) a risk assessment required to be carried out by Part 3, or
>
> (c) an assessment required to be carried out by section 14,
>
> involves a judgement by a provider about whether content is content of a particular kind.

Section 192(5) specifies that a provider is to have "reasonable grounds to infer that content is content of the kind in question", based on "all relevant information that is reasonably available to a provider" (s192(2)).

The most detailed Parliamentary discussion of this section – including cross-party concerns that the proposed illegal content judgement clause (then cl 173) would lead to over-takedown of content – was at Lords Report Stage on 17 July 2023. Other references to specific relevant points made by Lord Parkinson, the Government minister, are included below.

Ofcom's proposals

Volume 5 of the illegal harms consultation sets out their high-level approach to the Guidance ("How to judge if content is illegal or not?") and the draft Guidance itself is in annex 10 ("Online safety guidance on judgement for illegal content")

Safety by Design

The safety-by-design approach is central to the regime (s 1(3)) and should influence the implementation of both the illegal content safety duties and the children's safety duties; Ofcom will be consulting on the latter in phase 2 later this year. As set out in section 10,

there are a range of duties applying to illegal content (notably a general duty to mitigate) and some further duties applying to priority illegal content. These could be ex ante measures – for example, design choices (e.g. increased friction; approach to weighting of recommendation tools and revenue sharing policies), proactive measures (e.g. chatbot interventions to reduce racist messages) or ex post measures such as content curation and/or moderation (e.g. systems for downranking, takedown or account suspension). All the OSA duties relate to the design of the service (broadly understood) or its operation, and not to individual items of content (ss 10 and 27). **This is not fully reflected in Ofcom's discussion of the meaning of illegal content in Volume 5 or the approach taken in the draft Guidance (Annex 10) – even when taking into account the constraints of the Act.**

Ofcom's discussion focuses on individual items of content – to the point of saying in the draft Guidance that the obligation is to take content down (Annex 10, para A1.14), rather than - for user-to-user services - to operate a proportionate system designed to have that effect. It is unclear in the main how this becomes relevant to search. Of course, the operation of a content moderation system does imply the application of the rules to individual items of content, but that is not the primary obligation in the Act. Lord Parkinson made this clear at Lords Report stage in the discussion on this clause.

> "My Lords, I start by saying that accurate systems and processes for content moderation are crucial to the workability of this Bill and keeping users safe from harm." (Hansard July 17 2023 col 2141)

And:

> "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." (Hansard July 17 2023 col 2143)

Earlier in the Lords, at Committee stage, Lord Parkinson also said:

> "To be clear, the duty requires platforms to put in place proportionate systems and processes designed to prevent users encountering content. I draw my noble friend's attention to the focus on systems and processes in that. This requires platforms to design their services to achieve the outcome of preventing users

> encountering such content. That could include upstream design measures, as well as content identification measures, once content appears on a service." (Hansard 27 April 2023 col 1359)

Ofcom could have chosen a different approach – one which fits with the systems approach – within the terms of the Act. The definition of illegal content in section 59 does not specify whether the requirements of the offence are to be defined in the abstract or in the applied context. When considering the application of the rules in a system (e.g. in the take-down context where an individual item of content is in issue) it might be relevant to assess how the behaviours required by the criminal offence may map on to those of the user. When we are looking at the design of the systems and processes, however, considering the offence more generically makes more sense. The effect of the Act being drafted in systems language means not only that understandings of illegal content must be considered at scale but also before those items of content have come into being. The recognition in the draft Guidance that services may be dealing with content "in bulk" (Annex 10, para A1.15) is not quite the same point.

Moreover, the word "content" is ambiguous. In line with usual statutory interpretation, the singular includes the plural, but could also extend to types of content. Section 192(4)(a) distinguishes between content and kinds of content suggesting that both approaches should be covered. The Guidance has considered when inferences could be made, focussing on an item-by-item approach. It should also consider what the signals for inference (the "reasonable grounds" in s 192(5) and (6)) about the mental element of crimes under s 192 are in relation to systems design. These cannot be the same for systems as in individual items of content (where Ofcom suggests that decisions should be made on a case-by-case basis (Vol 5, para 26.24, 26.82) – which would in any event be hard to scale, even ex post).

See for example Lord Parkinson's statement at Committee stage in the House of Lords:

> "Companies will need to ensure that they have effective systems to enable them to check the broader context relating to content when deciding whether or not to remove it. This will provide greater certainty about the standard to be applied by providers when assessing content, including judgements about whether or not content is illegal. We think that protects against over-removal by making it clear that platforms are not required to remove content merely on the suspicion of it

being illegal. Beyond that, the framework also contains provisions about how companies' systems and processes should approach questions of mental states and defences when considering whether or not content is an offence in the scope of the Bill." (Hansard 27 April 2023 col 1358)

Significantly, the Guidance emphasises that inference is based on the substance of individual items of content in relation to all priority offences. While this is a starting point, the service could have other sources of information on which to base its judgement. While a broader range of sources of information are noted in the Guidance (Annex 10, A1.66 and for examples see A6.11, A6.24 and A 6.37 in re fraud), these are not consistently considered. Contextual information is important when determining categories of content – for example, patterns of posting (e.g. frequency and timing of posts could be relevant for understanding harassment; cf Annex 10, A3.100-102), the existence of networks in addition to the content-based context of what was in the post before or after the impugned item (which Ofcom notes in some instances).

Of course, the precise significance of different types of contextual information may vary between offence type (for example, there is - as Ofcom notes - less to understand in the context of CSAM than threats). Nonetheless, Ofcom's focus in the Guidance seems to be on the context of the content itself (see e.g. examples in Annex 10, A 2.17, A6.65) rather than a wider range of metadata context information which may be of particular use when identifying categories of content and designing systems.

There is a significant gap in the Guidance here. While Ofcom rightly notes that in practice user-to user services' terms of service cover more content than that which would be defined as illegal content (on relationship between illegal content judgements and terms of service see 26.17 – 26.18 and 26.43 – and note that search engines do not have to have terms of service), the definition of illegal content is important beyond providing a floor for those terms of service. **Illegal content defines the scope of the regime as regards the illegal content duties.**

Volume 5 and the draft Guidance focus – as do Volume 4 and the draft Codes – on ex post measures, specifically takedown (see e.g. para 26.43 and in the draft Guidance that search can only ever look at individual items of content – Annex 10, para A1.16). They do not consider ex ante measures (which are not limited to proactive technologies within the meaning of s 231) nor safety by design, and so do not consider how measures which

are not targeted to particular content but are aimed at generally removing risk/improving safety (and thereby impact across a range of harms caused by different types of illegal content) will fit in. For example, recalibrating the weighting on recommender tools (perhaps in line with the adaptation of such tools under the DSA as suggested in recitals 87-89) or taking steps to deal with data voids.

These questions are important as it is on the design/operation of system that the service can satisfy its duties and not on the taking down (or not taking down) of specific items of content. So while the draft Guidance covers some of the ground, Ofcom should consider how to understand content by reference to systems, and make clear that the Guidance, as is, is not exhaustive in that regard.

Content not Conduct

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. The requirement for reasonable grounds to infer a criminal offence each time content is posted, as outlined in Vol 5 (para 26.44 et seq) and the draft Guidance, presents an overly restrictive interpretation of relevant content. Such a narrow perspective is not mandated by the language of section 59, which necessitates the existence of a link at some stage, rather than in relation to each individual user. The significance of this can be seen in the example given of the reposting of intimate images without consent – the re-post is still the content linked to the original offence, it has not changed its nature. Contrary to the views expressed in Annex 10, para A1.59, there is a difference between the same content and altered content. There is no obligation in the Act to look at the mental state of each individual disseminator of the content. Moreover, this point needs to be understood against those made about the systems obligations, when design choices are made in relation to types of content rather than specific items.

Burden of Proof in a Civil Regime

The Act introduces a civil regime not a criminal one. The Consultation recognises that "'Reasonable grounds to infer' is not a criminal threshold", and further notes that this test is the relevant test rather than beyond reasonable doubt (see Vol 5, para 26.14). Given that the regime is a civil regime, rather than a criminal one, this threshold should be understood against the civil burden of proof – that is on the balance of probabilities.

This means the threshold for proof is lower both as to the types of evidence considered to give rise to an inference and the amount of evidence required. There is also the question of how to approach inference in the absence of evidence being reasonably available – to what extent (especially with content implicated in more serious crimes) should such an inference be made; it may be that there is greater scope for some offences (e.g. those with low mental element thresholds) than where there are more specific requirements.

In this, we should note that there is a wider body of potentially relevant information – the wider context Lord Parkinson referred to. So, for example, where we have evidence about widespread negative impact of a behaviour (e.g. cyber-flashing), with little in the way of countervailing interests (contrast for example the difficulties around the suicide offences), we could infer that the mental element was met. Evidence, albeit limited, indicates that a proportion of men know that the images cause distress. Moreover, Ofcom's own evidence gathering, set out in Volume 2, says that "Cyberflashing is not a product of technology and online behaviour alone; it is a manifestation of existing patterns of sexual violence and abuse. McGlynn argues that cyberflashing should be understood as part of a continuum of sexual violence. As with all forms of sexual violence, perpetrators of this abuse are motived by a desire to exert power, and victims and survivors experience feelings of fright and vulnerability." (Vol 2, 6S.19).

Given this understanding of the nature, extent and severity of the harm, is it not reasonable to infer on the balance of probabilities that the content is linked to criminal behaviour? In the context of articles used for fraud, Ofcom proposed "when considering the user's state of mind, services should ask themselves whether there is any possible use of the article concerned which is not for fraud" (Annex 10, A6.66). Yet for cyberflashing Ofcom suggests – without explaining why – that it would be hard to infer the mental element (Annex 10, A 10.43). The approach Ofcom has taken here is unnecessarily restrictive – especially as Ofcom has in relation to terrorism suggested that the threshold of recklessness is reasonably easy to infer (Annex 10 A2.55, A 2.69). As a consequence, Ofcom fails to deal with this harm. It also raises the question as to whether the standards of proof in the Guidance are consistently those of the civil regime, or whether in some instances, a narrow approach has been adopted.

Moreover, the difficulties in these areas are compounded because Ofcom has considered inference in the face of a lack of evidence in respect of a moderator on a

case-by-case basis (Vol 5, para 26.82). As noted, signals on which inferences may be made, may need to be understood differently than in the context of a case-by-case analysis.

Missing Offences

Ofcom is right in asserting that identifying the most serious or most specific priority offence is not the most effective way to think about how the regime works; for the purposes of the regime, it is sufficient if any priority offence is triggered and so the broader priority offences are the most significant when it comes to triggering the regime. So, when an offence (and the consultation gives the example of racial hatred) is committed, for the purposes of applicability of the illegal content duties and enforcement it does not matter whether it is the aggravated offence or the base offence.

Against this recognition, it is unfortunate that Ofcom has not considered any of the existing non-priority offences, specifically s 127(1) Communications (which unlike s 127(2) Communications Act, has not been repealed) or the Obscene Publications Act 1959 (listed as priority in Sch 6 in relation to those offences only). Much content falling out of more specific offences will be caught by the Obscene Publications Act or by s 127(1), and therefore some safety duties would apply, notably the base level of mitigation (s 10(2)(c)) and having a system to take content down (s 10(3)(b)). The existence of these offences should be flagged so that they are not forgotten or overlooked, especially as Ofcom has suggested it is not proportionate for providers to anticipate all non-priority offences (Vol 5, para 26.70) and that (in relation to terrorism offences) the giving of guidance in relation to some offences and not others is to suggest to providers where they should focus their attention (Vol 5, para 26.64). This approach makes sense where an offence is unlikely to occur; much less so where there are offences which are quite likely to be relevant, as is the case with the two offences here. Moreover, the selection of the non-priority offences in respect of which guidance is given is not based on the likelihood of them being relevant, but on their newness (Vol 5, para 26.72).

The Impact of Rights

It should be noted that Ofcom has an obligation to take into account fundamental rights, noted in para 26.8 and reflecting the requirements of the Act, and this has weighed towards a narrow interpretation of illegal content. However, the terms of the act cannot

remove the Ofcom's obligations in relation to other fundamental rights. While intrusions into Article 10 must be carefully considered, three counter points should be noted.

First, the speakers are *not* being *criminalised* by the application of the regime – this means there is a lesser intrusion into their rights than there would be were criminal penalties to be imposed. Even the takedown of content for legitimate reasons is a more proportionate response than the imposition of a criminal penalty. Indeed, takedown has been found to be a proportionate response in relation to civil actions; account removal – which has a greater impact on the user's speech rights - has in the case of persistent violation, been found appropriate in a regulatory regime (see *NIT S.R.L. v. the Republic of Moldova* (28470/12)).

Secondly, survivors of online harms have rights too, which should be taken into account, as is discussed here, and this Ofcom has completely failed to do – especially as regards the well-documented silencing effect some content has on others, especially those in minoritised groups. (See analysis on Ofcom's approach to human rights in the illegal harms consultation here.)

Further, while this point may be implicit in some of Ofcom's analysis, it should be expressly recognised that some content is likely to be more worthy of protection than others – and that this affects the impact of freedom of expression concerns on scope of offence. While it is possible that abusive speech could in some instances be considered to be political speech which attracts significant protection from Article 10 (see e.g. *In re S (FC) (a child)* [2004] UKHL7, albeit in the context of a civil law claim), it is hard to think that sharing deepfake porn would do so (even though it formally falls within Article 10).

RECOMMENDATION

We would urge Ofcom to review the approach it has proposed in the light of the analysis above. We recommend that Ofcom consider how to revise the Guidance before it is published to address the risks that the current focus on a piece-by-piece approach to content will have for the effectiveness of the regulatory regime as a whole. At a minimum, an additional focus on the application of systemic and by-design measures – as provided for in the Act - should be added to the Guidance to ensure providers can apply it at scale.