ONLINE SAFETY ACT
NETWORK

EVIDENCE FOR THE SCIENCE, INNOVATION AND TECHNOLOGY COMMITTEE INQUIRY INTO SOCIAL MEDIA, MISINFORMATION AND HARMFUL ALGORITHMS

We welcome this new inquiry by the Science, Innovation and Technology Committee and are pleased to submit this evidence to inform its deliberations.

The Online Safety Act Network brings together over 70 civil society organisations, campaigners, academics and advocates with an interest in the implementation of the Online Safety Act 2023 (OSA). More details about our work are here. The Network continues the work carried out by Professor Lorna Woods OBE, William Perrin OBE and Maeve Walsh at Carnegie UK during the passage of the Online Safety Bill: Professor Woods' proposal for a “duty of care” to address online harm reduction formed the basis of the OSA; and the Carnegie team supported Members, Peers and Select Committees during the Bill's passage, gave evidence to Parliamentary inquiries and Bill Committees and were acknowledged by Parliamentarians in both Houses for their contribution.[1]

Our primary evidence to this inquiry is new analysis by Professor Woods which is relevant to both question 3, on the role that social media platforms played in the riots in summer 2024, and question 4 on the UK's regulatory and legislative framework and the OSA specifically. In the final section, we suggest a number of recommendations for Government and for Ofcom that flow from this analysis. We also provide supplementary material on the background to the OSA which may be helpful to the Committee. This includes:

- Our OSA 2023 explainer: here.
- Our user-to-user illegal harms duties explainer: here, and attached at annex A.
- Our initial reaction to the limits of the Online Safety Act in dealing with mis- and disinformation in the context of the post-Southport riots: here and attached at annex B.

[1] See, for example, Lord Clement-Jones: “I also pay my own tribute to Carnegie UK, especially Will Perrin, Maeve Walsh and Professor Lorna Woods, for having the vision five years ago as to what was possible around the construction of a duty of care and for being by our side throughout the creation of this Bill'. (Lords Hansard: 6 September 2023, Column 470)

- A paper by Professor Lorna Woods on safety by design: here and at annex C.

- Our summary response to Ofcom's illegal harms consultation: here; and attached at annex D. An annex to our full response is also attached at annex G.

- Our analysis on Ofcom's draft Illegal Content Judgements Guidance: here; and attached at annex E.

- A copy of a paper written by Prof Lorna Woods in 2019 which looks at the "duty of care" approach to online harm reduction in relation to fundamental rights, including freedom of expression, at annex F.

Question 3: What role did social media algorithms play in the riots that took place in the UK in summer 2024?

Question 4a: How effective is the UK's regulatory and legislative framework on tackling these issues?

Question 4b: How effective will the Online Safety Act be in combatting harmful social media content?

Prof Woods has provided the new analysis below which covers these interconnected issues.

The Southport riots involved two broad categories of content (clearly illegal content on the one hand; and misleading and inaccurate content on the other) which can be used as models to assess how the Online Safety Act (OSA) regime might work when it is enforceable and to demonstrate the limits to that regime. The Southport riots also raise questions about the relationship between social media and content producers of different kinds: traditional media; outlets masquerading as news outlets; figures with large followings (including influencers, commentators and celebrities); and figures with small followings/networks.

The Online Safety Act's Approach to Content

The OSA applies different rules to regulated content depending on whether the content is "illegal content" or "content harmful to children"; further rules apply (obligation to provide user empowerment tools; obligation to enforce terms of service) which are not limited to these two categories of content; they are, however, limited in terms of the services to which they apply (user-to-user Cat 1 services).

Within the two categories of content there are subsets of content called "priority content" (and in relation to content harmful to children, "primary priority content"). There are more specific duties in relation to both sets of priority content. The obligations are phrased differently for user-to-user services (essentially social media) and search, with the obligations on search being slightly less onerous than those applying to user-to-user services. A table setting out the specific duties can be found here. Our explainer - here and at Annex A - provides more detail on how the illegal content duties and risk assessments work and what enforcement measures Ofcom has at its disposal.

Hateful Content

A significant number of people were charged with criminal offences for social media posts. The majority of these were clearly hate speech or public order offences. For example, Jordan Parlour was given a 20-month custodial sentence for posting written material intending to stir up racial hatred (section 19(1) of the Public Order Act 1986, as amended). Among other posts, he posted on Facebook in early August, in relation to a hotel housing asylum seekers, that:

> "every man and his dog should smash [the] f*** out of Britannia hotel (in Leeds)"

In sentencing remarks (Rex v Jordan Parlour), the judge noted the role of the platform, Facebook, in allowing his posts to reach a larger audience (p 3):

> “The initial post received 6 ‘likes’, however it was sent to your 1500 Facebook friends and because of your lack of privacy settings will have been forwarded to friends of your friends.
>
> The messages were therefore spread widely which was plainly your intention.”

Tyler Kay was similarly charged with the offence of publishing written material intended to stir up racial hatred (section 19(1) of the Public Order Act 1986, as amended). According to the sentencing remarks (Rex v Tyler Kay), he made a public post on X stating:

> “mass deportation now, set fire to all the fucking hotels full of the bastards for all I care… if that makes me racist, so be it”.

The post added the hashtags: #standwithlucyconnolly #fucknorthamptonshirepolice #conservative #farageriots #riotsuk #northampton. The hashtag “#standwithlucyconnolly” seems to refer to another person charged and convicted for similar posts. Other posts included screenshots of posts that had been removed but which incited violence against immigration solicitors, also with numerous hashtags.

The judge noted:

> “The nature of the social media platform means the posts could have been viewed by any member of the public. The captured posts show views ranging from the low hundreds up to 3457 for the post referring to setting fire to hotels which amounts to widespread dissemination.”

A third example of the same offence is found in Rex v Connolly. In that case, Lucy Connolly posted on 29th July 2024 on X as follows:

> “Mass deportation now, set fire to all the fucking hotels full of the bastards for all I care, while you’re at it take the treacherous government and politicians with them. I feel physically sick knowing what these families will now have to endure. If that makes me racist so be it”

The tweet remained available for at least three and a half hours and was, according to the sentencing remarks, widely read – it was viewed 310,000 times with 940 reposts, 58 quotes and 113 bookmarks. She had also posted other racist comments. In sentencing, the judge noted the intention to achieve widespread dissemination of her remarks.

News media report other cases, including one involving the posting of AI-generated images. Most seem to be offences for stirring up racial hatred, a priority offence for the purposes of OSA.

While many cases involve X, this is not the only platform: Parlour posted on Facebook and in Rex v James Aspin, TikTok was used (offence of distributing a recording intending to stir up religious hatred). Again, in Aspin, the fact that the account was unrestricted was noted. While in many cases the defendant pleaded guilty, a number of jury trials found defendants not guilty.

These offences are priority illegal offences and the cases listed are clear examples of those offences, especially given the context of the Southport riots; whether all hateful content would be equally clear is another question. Meta’s Oversight Board has also opened an investigation into Meta’s decision to leave up certain posts which were reported to them for violating Meta’s policies on hate speech or violence and incitement. The investigation focuses on three posts: one referred to migrants as terrorists and called for more mosques to be smashed and buildings to be set on fire where “scum are living” (which Meta subsequently confirmed had been left up in error); another contained AI-generated images of Muslim men being chased and included the hashtag “EnoughIsEnough”; the third is a repost of another likely AI-generated image of four Muslim men, one of whom waves a knife, chasing a blond toddler in a Union Jack t-shirt - overhead a plane flies towards Big Ben. Meta confirmed to its Oversight Board that these two posts were correctly left up.

Some other offences will in general be more difficult, notably the s 179 OSA offence concerning false communications. A commentator who shared inaccurate information (Spofforth) was ultimately not charged. This case raises questions as to how easy it would be for services to identify content meeting the criminal threshold for this offence. As a background to the difficulties relating to the definition of illegal content, the Committee are invited to read our detailed analysis on how Ofcom proposes that services should judge whether content meets the illegal content threshold - provided here and at annex E. Note this was based on the draft consultation documents; while some changes have been made there does not seem to have been a fundamental revision in approach in Ofcom's final Statement which, at 457-pages long, we have not had time to analyse in the two days between its publication and the Committee's call for evidence deadline.

The Illegal Content Duties

The offences trigger the priority illegal content duties. Consequently, not only should platforms be seeking to mitigate harms caused and to have a system in place that allows the swift removal of illegal content on notification of it, but they should take proportionate design or operation measures to prevent users from encountering it, effectively to mitigate the risk of the service being used to facilitate an offence and to minimise the length of time such content is available. These latter obligations tend to be viewed as proactive; they can apply at a range of points in the content distribution chain and not just at the point of content moderation and take down.

In earlier work, such as the paper attached at annex F, Professor Woods has described the communication distribution chain as comprising the following stages:

- **User onboarding and content creation** - e.g. ease of account creation (including anonymity; swift sign up processes) and "disposable" or bot accounts; what augmented reality filters are available; privacy settings; friction nudges about standards; metrics and monetisation.

- **Content discovery** - eg recommender tools; accounts to follow; hashtags; and trending topics.

- **User engagement and reaction** - eg likes/upvotes; user empowerment and curation tools; complaints.

- **Service response: moderation and appeals** - eg response to complaints and own initiative (removal and down ranking); user appeals; review of accounts.

Ofcom's recently published Code on Illegal Content[2] focuses on companies' need to set up internal governance mechanisms and content moderation processes. The Online Safety Act distinguished between user-to-user and search. As regards user-to-user, the Act does not provide an obligation on companies to respond to notifications in a particular timescale, but rather to operate a system that allows them to respond swiftly to notifications - in this there are questions about the trade off between speed of response and accuracy of response. This approach means that a service can satisfy the OSA obligation even if they do not take all content

[2] At the time of writing, Ofcom's final code on illegal content and accompanying documentation have just been published (16/12/24); we link to it above. Our analysis is based on our reading of the versions of the documents published for consultation; while we have not had time to go through the detail of the new, lengthy documents, Ofcom has not made many material changes to the content so - except where clearly noted - our analysis stands.

down. The question is, at what sort of level of missed content does a company fail to satisfy this requirement? This obligation should be understood against the other obligations around process which require services to have complaints and appeals processes to deal with false positives and false negatives (see s 21).

There are no hard obligations on this point in the Act or in Ofcom's Code, rather the Code allows the providers some freedom to set their own performance targets, and to choose how to prioritise content. Presumably Ofcom will assess whether these are appropriate to satisfy the obligation in the OSA - one would assume that a slow response with a significant error rate, at the very least, should not be acceptable. It would seem sensible to have a system in place that kicks in in circumstances such as the Southport riots, so that relevant content can be dealt with swiftly. The Code does not include that as a specific requirement, nor are there minimum thresholds for resourcing of moderation functions.

Search services have no obligation to run a take down system but rather to engage systems to minimise the risks of individuals encountering illegal content that the provider knows about (s 27(3)). Ofcom suggests appropriate moderation for search means that the content no longer appears in the search results presented to UK users or that the content is deprioritised (Ofcom Statement, para 3.60). This means a search service could satisfy its duties here even if it still presents illegal content in search results. Where the search results can only return illegal content, it seems that Ofcom accepts that the service provider can return those results (Ofcom Statement, para 3.65).

Ofcom's suggestions for earlier up the communication distribution chain are more limited, despite the OSA's focus on the full communication chain and its concern about virality. The main focus for Ofcom's recommendations beyond moderation relates to CSAM and terrorism. Additionally, for higher-risk services, Ofcom recommends safety testing for recommender tools and providing enhanced user controls. Recommender tools, as Ofcom's Risk Register recognises, can play a role in promoting hateful content as it might promote user engagement, even if in the form of outrage (Ofcom Risk Register - 3.68-3.71). Ensuring that recommender tools do not have this unintended consequence by testing them should be helpful.

While user empowerment tools can be helpful in some circumstances, they would not seem to be helpful here as some people will be seeking out like-minded folk – a tendency Ofcom notes in its draft Risk Register (para 3.27-3.29). Some functionalities identified in Ofcom's draft Risk Register as presenting a potential risk in this space are not addressed in Ofcom's Code of Practice[3] – for example anonymous accounts (though anonymous accounts also provide significant benefits) and ease of reinstating an account even with a very similar name after having been suspended. Hashtags, as well as a means to identify content and to promote it, can

[3] We attach at Annex G a table that shows the disconnect between Ofcom's register of risks and the measures it proposes to deal with these risks in the illegal harms codes. While this is based on the consultation documents, there have been no new measures added by Ofcom in its final version so the analysis here still stands.

lead to the creation of communities of like-minded individuals (Ofcom Risk register para 3.60). The sentencing remarks (above) also note features such as privacy settings and hashtags that play a role in spreading hate, even for users without significant followings.

Inaccurate and Misleading Information

Illegal Content

Outside fraud offences, there are two main offences relevant to inaccurate and misleading information: the foreign interference offence (which is a priority offence) found in s 13 National Security Act 2023 and s 179 OSA. The foreign interference offence did not seem to come into play on the facts at the time, but a small number of cases arose in relation to s 179. *Spofforth* concerned the circulation of false information on X that wrongly identified the suspect in the deadly Southport knife attack as being Muslim and an immigrant. The post was deleted shortly after posting and Spofforth claimed she thought the information was true. S 179 relates only to knowingly false information. Consequently, the prosecution did not proceed. Conversely, Dimitrie Stoica pleaded guilty to the offence; he falsely claimed on a live-streamed TikTok video (to 700 followers) that he was "running for his life" from rioters in Derby.

The issue of identifying whether s 179 is satisfied for the purposes of OSA gives rise to some difficulty; this is - as noted above and at Annex E - part of a more general problem with the definition of illegal content. Ofcom notes that a service will not always be in a position to judge and that unless there are grounds to infer that it is, the content cannot be judged to be illegal (Ofcom Illegal Content Judgments Guidance ("ICJG"), para A13.24)[4]; Ofcom does note that services may have grounds to infer on receipt of information from law enforcement or by reference to a court order. Ofcom's analysis very much focuses on an analysis of the content at the time of moderation and not how the earlier stages in the distribution chain might affect content (e.g. monetisation giving rise to click bait and rage bait content, or high-risk stunts). In this context, Ofcom does not consider how its general approach on information available to service providers (the content itself; complaints information; user profile information and activity and published information from credible and relevant sources) affects the ability of services to make an inference about likely knowledge of falsity.

Virality and the Criminal Law

Ofcom's approach in its ICJG assumes that the language used in s 59, which describes "illegal content" for the purpose of triggering the illegal content duties, requires a criminal offence to happen each time content is posted (though this is not expressed on the face of the OSA). Ofcom argues (Vol 5, para 26.45):

[4] Our analysis here is based on the draft guidance published for consultation in November 2023

> “when a piece of content has been shared, forwarded or reposted by a new user, a service should treat this as a new piece of content for the purpose of an illegal content judgement”.

Consequently, an item of content could flicker in and out of the regime depending on who posts or shares it; this interpretation is problematic from the perspective of tackling viral content. Focusing on requiring all elements of a criminal offence and in relation to individual posting means the regime may be incapable of dealing with viral content in relation to a large range of offences. While Ofcom has changed its guidance in relation to some offences on this point (eg non-consensual intimate image abuse) it has not dealt with the issue generally. In the case of s 179 OSA, as it is shared through networks the knowledge of its source, context and consequently truth or otherwise may be lost; this does not mean that the content ceases to be problematic from the perspective of the regulatory regime. The argument that this stops people from ‘calling out’ content by re-circulating it is not convincing as a counter argument; it merely gives the illegal content more publicity (see concerns expressed by e.g. Phillips (2018) in relation to how the media reports on extremism) and encourages a vigilante mindset. While there is rightly a concern about over-criminalisation of people, it is important to remember that we are talking here about the threshold for the applicability of a civil regime, which does not have the severe consequences for individuals that the criminal law has.

Content Harmful to Children

The categories of content harmful to children do not encounter the same problems around mental elements as we see in the criminal law; the focus in relation to this category of content is on the impact on individuals. Moreover, the focus on impact allows the effect of cumulative exposure to content that taken individually is not that harmful to be taken into account (this issue could be problematic in determining whether cumulative sub-criminal posts taken together cross the threshold for illegal content – it would seem possible only where the offence itself allowed for this (eg harassment)). Some of the content could constitute priority content harmful to children, for example, abusive content targeting those with a protected characteristic or content which incites hate, bullying content and content which encourages an act of serious violence against a person (s 62 OSA).

Services likely to be accessed by children would then have to satisfy the children’s safety duties in respect of that content. A key mechanism for doing so appears to be age-gating the content or the service, which does nothing to tackle the circulation of the content more generally. It is also noteworthy that misinformation is not per se content harmful to children – only misinformation that is harmful would be caught. Suggestions that some categories of misinformation that were harmful would be included in the Act (eg health misinformation) did not come to fruition; potentially these other forms of disinformation could be non-designated content harmful to children when “it presents a material risk of significant harm to an appreciable number of children in the United Kingdom” (s 60(2)(c)).

Sub-criminal Inaccurate or Misleading Information

For non-criminal content that is not "content harmful to children", the main mechanism to deal with it is through the provisions on terms of service. Section s 72(3) of the OSA in effect provides that Category 1 services must apply their terms of service and do so equally. Some services provide prohibitions on hate speech and on misinformation - though how these prohibitions are defined is the choice of the service. The DSIT Secretary of State has recently laid the regulations required to set the categorisation thresholds based on advice from Ofcom and Ofcom will in due course publish its register of categorised services. This obligation will not apply to all user-to-user services and does not apply to any search services. Furthermore, services may only act against content if their terms of service provide so (s 71). Notably, **the Act provides for no minimum content in relation to terms of service, so it is possible that some services might not deal with issues of misinformation at all**. It also does not stop providers from reducing the level of protection (as, for example, X did). Ofcom could do nothing in this scenario.

Crisis Protocols

Unlike the Digital Services Act, the OSA does not require platforms to implement triage systems or crisis protocols though – as noted – they would seem a sensible internal mechanism for services to have in place to deal with inevitable crises and disasters. The OSA does have a provision dealing with "special circumstances" (s 175). This is an odd provision as the motivating initiative lies with the Secretary of State and relates either to the health or safety of the public (which would seemingly cover riots) or to national security. The Secretary of State may give a direction to Ofcom to do one of two things:

- Use their media literacy powers to give priority to objectives specified in the direction; and/or
- give a "public statement notice" to a service provider, which requires the provider to make public the steps they are taking to tackle the threat.

The provision does not allow Ofcom to tell service providers what to do; nor does it allow Ofcom (or the Government) to require particular pieces of content to be taken down. Moreover, the service provider cannot act otherwise than as specified in its terms of service (s 71 OSA). Of course, any direct involvement by public bodies in content being shared on communications networks needs careful oversight, if acceptable at all, to ensure protection for freedom of expression. We note that Ofcom has announced it will consult on crisis response protocols for emergency events (such as last summer's riots) as part of its second consultation on the illegal harms codes.

The Wider Picture

Part of the story relating to the riots in the summer is about the spread of inaccurate information hosted as follows: on sites which ape the look of online media outlets but are farming clicks; by figures with significant followings (including celebrities, influencers and commentators or various sorts); and by bot networks. AI "slop" will arguably make the position worse. Finally, even those with limited followings may have an impact when their content gets a wider reach in relation to a specific event (here we can see the role of hashtags in combination with larger voices - see for example the use of hashtags in the *Kay* case (above) - especially when these reflect slogans, eg: "Stop the Boats," "Enough is Enough"). This pattern can be seen in the Southport riots as a number of analyses have detailed (see e.g. ISD here), where content is picked up either by so-called news outlets, or by commentators with large followings and spread even after the original post is taken down. It may be that those who have a strong role in opinion formation, or accounts and creators which claim to be media outlets, should be subject to some sort of quality standards (which may then help social media and search, and even advertisers, distinguish between different qualities of speaker).

It is notable that influencers, who may spread content to large audiences, do not tend to check for accuracy; how popular a post is seems more pertinent to them. UNESCO has suggested the need to train influencers, though other studies suggest that it is "heavy users" who tend to spread disinformation rather than necessarily influencers (the two groups may of course overlap); posting and sharing eye-catching information seems to be habit forming, or money-making. In terms of those taking advantage of events and disasters, it may be that crisis protocols would be helpful (though not a substitute for policies addressing root causes of the problem).

At the moment, the traditional media are subject to standards around accuracy (though that pertaining to the press is much weaker than the broadcast regime). It could be that some audio-visual content on social media platforms already falls within the definition of "on demand programme service" (ODPS) for the purpose of the video-on-demand regime. Historically, this had very minimal standards (though it did prohibit hate speech). The Media Act 2024 brings in more extensive rules around accuracy and impartiality but only for a sub-category of ODPS. While the provisions introducing the expanded regime have been commenced, the details of the regime remain to be finalised. Specifically, the sub-category of services to which the extended rules should apply is to be determined by the Secretary of State in secondary legislation; the regulations have not yet been laid. The Government recently wrote to Ofcom asking them to start the preparatory work, Ofcom's roadmap suggests that it will be consulting on a code in Spring 2025. The Media Act also updates the rules for prominence of public service broadcasting content in relation to certain sorts of services. It was envisaged that these rules should apply principally to smart TVs and the like - the Internet Television Equipment Regulations 2024 apply the rules to smart televisions and streaming devices, but there is a question as to whether the rules should be expanded to other content recommendation tools and apps. If such a proposal were to be taken forward, careful consideration would be needed as to which services should benefit from any expanded prominence rules.

As regards "disinformation for hire" or computational propaganda as a service, it may be that these sorts of activities should be limited by law – perhaps through the regulation of ad agencies and PR firms, or by the prohibition of certain sorts of activity. Other commentators (Lee, 2020) have argued that PR normalises "organised lying" and plays a role in the creation and dissemination of disinformation (see also Grohmann and Corpus Ong (2024).

Question 4 c: What role do Ofcom and the National Security Online Information Team play in preventing the spread of harmful and false content online?

As we have noted in our analysis of the limits of the OSA (link above, and at annex B), Ofcom's role is prescribed by the duties that it is required to enforce within the Act. The letter from its Chief Executive to the Secretary of State in October (here) sets out how far they feel they can act, once the Act is fully in force. For Ofcom to go further in relation to preventing the spread of harmful and false content online, they could take more of a "safety by design" approach - as we have argued in our attached paper; however to fully address the challenges that arise to social cohesion, public order and democracy, the Act would need to be amended. We have set out some recommendations in that regard, related to Prof Woods analysis, below.

It is right that the Committee ask about the National Security Online information Team but this question is best addressed to the Government. Successive Governments have refused to reveal any details about the work of this team, or its predecessor organisation, the Counter-Disinformation Unit. Here is a recent answer to a question by the DSIT Minister Feryal Clark, which refers to the riots.

We have argued for many years - when previously working with Carnegie UK - that the Counter-Disinformation Unit needed to be put on a statutory footing, with appropriate oversight and transparency. The same remains true of the new unit. Here is an extract from our submission to the Joint Committee on the Draft Online Safety Bill in September 2021 (the full submission, which also deals with the lack of measures to address mis- and disinformation in the Bill, can be found here)

> "The draft OSB … should be an opportunity to lock significant platforms into a risk assessment mechanism for threats to security from mis- and disinformation under regulatory supervision, with appropriate transparency to Parliament. The draft OSB could also formalise and make more transparent the manner in which the UK public sector communicates threat assessment to platforms through the operation of the Counter Disinformation Cell in DCMS. The Cell should be put on a formal statutory footing with an obligation to report to Parliament and to include OFCOM in its work. On societal harms more broadly, we are concerned that the limitation of harms to

> individuals will not help the regime tackle issues such as high levels of misogyny and racism on a service which might undermine social cohesion, and indeed then feed back into harms to individuals. An avalanche of hateful speech in a public forum may have a greater effect on society than the sum of harms to individuals against whom it is directed. … The proliferation of misinformation and disinformation also has a corrosive effect on the country's "epistemic security", on people's ability to access and identify reliable information across a range of issues."

Which bodies should be held accountable for the spread of misinformation, disinformation and harmful content as a result of social media and search engines' use of algorithms and AI?

The regulated services can use AI in a number of ways, for example: augmented reality filters, or the integration of AI content creation tools more generally; nudges to remind users to comply with terms of service; trending topics; content recommendation tools; AI summaries of search terms; tools for identifying and removing problem content; chatbots for engaging with users. As a general principle, the use of AI whether as part of the operation of the service or in terms of users' use of AI should fall within the OSA. Ofcom has made this point. In this context, safety by design remains important, especially as regards AI tools that are integrated into the service. These should be fit for purpose. Services should also review their monetisation policies; we noted above how "AI slop" was being encouraged through the various platforms' monetisation policies.

While the responsibility and accountability for the spread of misinformation and disinformation as a result of social media and search engines' use of algorithms and AI should lie with the platforms themselves, the Act does not include misinformation or disinformation (unless and to the extent that it also constitutes illegal content or content harmful to children) as a type of content in regards to which services have any duties at all. So while the design of the service and choice of tools incentivises and enables bad actors, or those who are misled, to spread mis/disinformation far and wide, there is a gap in the regime. At present, neither the OSA, nor the powers that flow to Ofcom as a result, are fit for purpose should there be another online crisis based on misinformation, such as the reaction to the Southport murders last summer.

Recommendations

Based on the analysis above, and our more detailed critiques of the limitations of the OSA and the proposed approach to implementation (to date) from Ofcom, we suggest that the Committee consider the following recommendations.

For Government

The Online Safety Act should be amended to:

1. Include in the illegal content risk assessment an assessment of a) the role that functionalities beyond recommender tools play into amplification and virality and b) the role that monetisation policies play, with mitigation steps taken for both. A general obligation on services to take mitigating steps to address all the risks identified in their risk assessment assessment should also be introduced.

2. Remove the "safe harbour" provision (section 49), along with the specificity on code measures (Schedule 4), to ensure that Ofcom can include more outcome-focused measures to address a wider set of risks in future iterations of its codes.

3. Include minimum standards for terms of service for category 1 companies that need to be enforced.

4. Include stronger crisis mechanisms in the Act, considering both crisis response - as seen in the international sphere already (e.g., GIFCT Content Incident Protocol or requirements of Article 36 Digital Services Act) - and crisis-specific risk assessment and mitigation processes

5. Mandate Ofcom to produce a code of practice on safety by design, to underpin the existing largely content moderation-focused codes.

Consideration should be given to other legislative and regulatory mechanisms, including:

6. Whether the Media Act's rules for prominence of public service broadcasting content should be expanded to other content recommendation tools.

7. Which video on demand services should be in Tier 1 services for the purposes of the new rules introduced by the Media Act

8. Regulating the industry and agencies responsible for "disinformation for hire" or computational propaganda as a service.

In addition:

9. The Government must, without delay, put the work of the National Security Online Information Team - and any other central government teams engaging with social media platforms on disinformation - on a statutory footing.

For Ofcom

Ofcom's next iteration of the illegal harms code should include:

1. A specific requirement for services to have a system in place that kicks in in circumstances such as the Southport riots, so that relevant content and networks can be dealt with swiftly.

2. Minimum thresholds for resourcing of moderation functions to be identified.

3. A requirement that category 1 companies must not dilute existing terms of service.

4. Measures to address networks of anonymous accounts and the ease of reinstating an account even with a very similar name after having been suspended.

5. Stronger outcome-based measures, based on safety-by-design principles, relating to the testing and design of services.

6. A requirement on platforms to address all the risks identified in their illegal harms risk assessments, whether or not there are specific measures assigned to them in the codes. The codes should also be clear that mitigation of risks is not just about takedown of content but may also be about changes to the amplification mechanisms on the service.

Ofcom's illegal content judgements guidance should:

7. Review how the criminal threshold is understood to allow it to fit better with a systems approach and ex ante design-based mitigations.
8. Clarify the position that once a piece of content has been deemed illegal it stays that way.

We hope that the Committee finds this evidence submission useful. Prof Woods is happy to speak further to the Committee or to provide oral evidence at one of its hearings, if helpful.

Online Safety Act Network

December 2024