SUBMISSION TO JOINT COMMITTEE ON THE NATIONAL SECURITY STRATEGY (JCNSS) INQUIRY INTO DEFENDING DEMOCRACY

---

1. We welcome the opportunity to contribute to the JCNSS inquiry into "Defending Democracy". The Online Safety Act Network has been set up to provide policy development and advice to civil society organisations and to convene discussions on priority issues arising from the Ofcom consultation programme on the Online Safety Act (OSA), essentially continuing the work that Carnegie UK commenced on the online harms agenda.

2. As a result, we have an interest in online threats to democracy, including mis- and disinformation, election interference and abuse and intimidation of elected officials, which are not adequately covered in the existing legislation. We provide material relating to the relevant topics from the Inquiry's scope below.

What are the objectives, working methods and resources of the Defending Democracy Taskforce? What has it achieved since its creation in November 2022?

3. This is a difficult question to answer. The work of the Defending Democracy Taskforce is not in the public domain. With the exception of announcements on its establishment and first meeting and some defensive pro-forma responses to Parliamentary Questions since then, there has been no update on its objectives, working methods or achievements.

4. This lack of accountability or opportunity for scrutiny makes it difficult to assess whether its work is scoped adequately to meet current and future threats for democracy or whether its engagements and interventions, for instance with online platforms, are appropriate and effective. The situation is comparable to that of the Government's Counter-Disinformation Unit and Rapid Rebuttal Unit, neither of which are open to public or Parliamentary scrutiny and for which the Government's stock response for many years was to say that providing details on their work would "give malign actors insight into the scale of our capabilities". We set out our concerns about these units' lack of accountability in analysis for Carnegie UK at the time of the Lords Second Reading of the Online Safety Bill in February 2023. We note that the Government belatedly published a "fact sheet" on the work of both units last June but the CDU came under renewed scrutiny in October last year in the context of mis/disinformation relating to the Israel/Palestine conflict. The relationship between the CDU and the newly enacted regulatory regime remains unclear.

5. We hope that the Committee's inquiry can provide some of the material it needs to answer these questions and that the Government is urged to be more transparent in future on the work of this taskforce, particularly in this critical year for democracy in the UK.

Is there more that the Defending Democracy Taskforce could do before upcoming elections to protect political parties, elected officials and core electoral infrastructure?

6. Professor Lorna Woods OBE - adviser to the Online Safety Act Network - and its Director, Maeve Walsh, previously worked with Carnegie UK on proposals that influenced the core concepts of the Online Safety Act in particular the use of a statutory duty of care. We extract below some of the key aspects of our analysis and advice on measures to improve the online safety of people taking part in UK elections; the full proposals, published in 2021, are here. William Perrin - our former colleague and Carnegie UK Trustee – has also submitted separately on this topic to the Committee.

7. When cross-examined on the subject of online intimidation and threats to elected officials by Kim Leadbeater MP during Public Bill Committee on the Online Safety Bill, William Perrin OBE said:

> *'We are sitting here having this discussion in a highly fortified, bomb-proof building surrounded by heavily armed police. I do not think any member of the public would begrudge Members of Parliament and the people who come here that sort of protection. We sometimes hear the argument that MPs should not be recognised as special or get special protection. I do not buy that; no one begrudges the security here. It is a simple step to ask platforms to do a risk assessment that involves potential victims of harm, and to publish it and have a dialogue with those who take part, to ensure that the platforms are safe places for democratic discussion.' Thursday 26 May 2022 Col122*

Elections - safer social media for candidates and others

8. The National Police Chiefs' Council, the Electoral Commission the public service broadcasters and Ofcom already work on electoral security and traditional media regulation around elections. The Online Safety Act (OSA) now provides a risk management regime for social media but deliberately has no express election safety measures that consider the high risk to candidates.

9. The terrible atmosphere around events in the Middle East has led to increased physical security measures for MPs; a comparable focus on increased safety online is within the gift of the online platforms. Online safety of MPs and candidates can be improved by targeting within the OSA risk management regime, drawing on traditional media regulation. OFCOM and elections regulators should be challenged to adapt the OSA regime (and point out gaps for legislative action) to produce an election safety model along the lines below. This could also form the basis of enhanced protection between elections. Backed up by the OSA sanctions regime, this simple model would be a radical step-up in platform security for candidates.

10. Large social media companies should be required to treat elections as “sensitive events” (the terminology used by Ofcom to describe elections in its assessment of harms arising from the foreign interference offence), assess the risks and then mitigate those risks under regulatory supervision from Ofcom the Electoral Commission and the National Police Chiefs’ Council.

11. Large user-to-user and search services (self-funded on the polluter pays principle) should each:

a. Appoint a public-facing Elections Risk Manager working in the UK, with experience of prior UK elections to own risks arising on the platform.

b. Assess the risks arising to likely candidates and administrators (in case of USA-style intimidation of the count) by:

i. Surveying a representative sample of those people for their perception of risks, with particular focus on minority and minoritized groups

ii. publishing a forward look of their plans for all elections in the year (like the BBC annual election guidelines) and

iii. consulting the NPCC electoral violence specialists, senior officers of political parties, the European Commission and Devolved Authorities.

c. Consult publicly on the risk assessment including OFCOM, NPCC, the European Commission, political parties and Devolved Administrations to ensure that the correct risks are covered.

d. Revise the assessment and under regulatory supervision deliver risk mitigation plans working with candidates etc. These should include:

i. A rapid-response, 24/7 facility for candidates etc to raise risks, with the facility’s performance standards supervised by the regulator.

ii. Enhanced ability for candidates to filter social media for threats, abuse, intimidation etc. according to their risk tolerance

e. After each major election, publish a lessons learned report.

What role are emerging technologies, such as generative AI, expected to play in upcoming elections?

12. Significant concern has been expressed about the impact of generative AI in particular on the information environment generally, but specifically in relation to elections. AI may not create entirely new risks but it amplifies existing problems. Put simply, it will be hard to identify

when material has been altered or even created from scratch to affect voter views of candidates and policies; research suggests that for many people it is hard to distinguish. It has been suggested that using AI makes posts that are part of influence operations more persuasive; indeed, using gen-AI bots may allow ongoing connections with target audiences to be maintained. Moreover, the sheer increase in quantity of manipulated content can drown out reliable sources. The mainstream generators have made the creations of content accessible and easy; social media and to some extent search, facilitate the spread of such material (though there will also be issues about the traditional media's role). While the proliferation of deepfakes may as yet be at an early stage, the Centre for Countering Digital Hate reported that the volume of deepfakes on X (formerly Twitter) has been rising on average 130% per month; how this might translate to other platforms is unknown. The World Economic Forum suggested that between 2019 and 2020 the number of deepfakes increased by 900% - and this was before the general release of ChatGPT. Europol has estimated that by 2026 approximately 90% of content could be synthetic. Note that a number of tech companies have signed up to a voluntary accord to develop a common approach. Given the nascent state of development, it is unknown what impact this will have, if any. The World Economic Forum identified misinformation/ disinformation (including increase in quantity) as one of the top 10 global risks for 2024. As the US Cybersecurity and Infrastructure Security Agency (CISA) noted, AI can impact the cyber-security of elections, for example via sophisticated social engineering campaigns, or through the use of voice cloning at gain access to sensitive election administration or security information. Concerns around information and security could lead to a fall in confidence in the election results themselves. There is increasing public concern about the existence of deepfakes and their impact.

What can be done to improve public awareness of disinformation, fraud, and technological interference such as that through AI or deep fakes?

13. It is unclear how mis and disinformation, once it has been created is best tackled. There seems little academic consensus on the effectiveness of de-bunking and pre-bunking, or how these techniques are best effected. Even fact-checking and the provision of "official" information can be seen as problematic where population groups are disaffected. Despite the lack of clear evidence (either about effectiveness or lack of it), the labelling of synthetic content identifying its fabricated nature sits within techniques typically used in traditional media to disclose sources. While there is an appeal from a tested mechanism, there are come caveats for the online context. Such labelling needs to be both visible and understandable. Once shared, labels/watermarks might be removed; some have expressed concerns that audiences might assume that anything not labelled is therefore true. More recent research aims to incorporate watermarks in synthetic media although there have been questions as to whether it is possible to irrefutably embed watermarks. Further work is needed on detection and on showing provenance; platforms which rely on self-declaration by users are unlikely to be doing enough unless they plan to consistently and effectively enforce terms of service. Transparency alone, however, seems unlikely to be sufficient. Stronger actions could be taken around the

distribution of content, particularly synthetic content, with the objective of reducing virality and allowing more reliable information (whatever that might mean) to surface.

How effective is the UK's legislative framework for defending democracy, including the new powers under the National Security Act 2023?

14. We set out here some detailed analysis on two aspects of the current legislative framework which we hope will be of interest to the Committee in its deliberations and development of recommendations. The analysis is in two parts:

a) regulation of online broadcast content

b) the limitations of the Online Safety Act

Regulation of online broadcast content

15. The broadcast media has been subject to stringent rules around news and current affairs reporting, as well as the distribution of harmful content (including that which incites violence or hatred). Less onerous rules apply to video on demand services ("on-demand programme services" (ODPS) in the language of the Communications Act 2003). Beyond ODPS, online content is not directly regulated (eg traditional blogs). Increasingly there are audiovisual channels which seek to produce partisan programming labelled as news/current affairs which seek to avoid the rules on accuracy and impartiality for example, by seeking to position themsevles as user generated content on a social media platform. (See Ofcom's recent ruling on GB News.) At the same time, people are increasingly accessing their news through social media. This raises questions as to how the various regimes fit together and whether there are gaps (even taking into account the extension of the video on demand regime envisaged by the Media Bill). Gaps will, of course, limit the ability of Ofcom, as the responsible regulator across the relevant regimes, to address in a consistent manner hatred, abusive or extremist content that might be designed to influence the democratic process and that is available to UK viewers, and open the regimes to a form of forum shopping, even before we get to questions around accuracy and impartiality.

16. The history of this is as follows: rules regulating video-on-demand (VOD) in the UK derive from the Audiovisual Media Services. The Directive (AVMSD) and came into force in their current form on 19 December 2009. (See guidance here). They imposed a limited range of content rules on providers established in the UK; the rules did not cover service providers established elsewhere. Many providers based in the UK, such as commercial porn services, came within the substantive VOD rules but seemingly objected to being regulated (although they would be subject to constraints off-line). The result was that the providers relocated from the UK to other places (e.g. Canada) and then transmitted back to the UK from there. Post-Brexit, the criteria for establishing jurisdiction were amended to clarify when an ODPS would fall under UK jurisdiction but did not change the geographic scope of the rules. The Media Bill seeks to amend both the content-based and jurisdictional rules.

17. Currently, to be an ODPS, and thus a regulated entity, a service must satisfy a cumulative list of criteria, set out in s. 368A of the Communications Act 2003. If a service meets the criteria, the service provider should notify Ofcom of that fact (failure to do so can incur a fine) and comply with the Ofcom rules for ODPS (which include but are not limited to content standards). The central element of the definition is that the service provides programmes consisting of images or legible text (with or without sound). Services must be on-demand; live streaming lies outside the regime. By contrast to the Online Safety Act where services host user-generated content, the service providers here have editorial responsibility for the content.

18. Ofcom gives examples but not an exhaustive definition of services that provide "programmes", including: a 'catch-up' service from a broadcaster; a television programme archive service; an on-demand film and television service, provided online by a person with "editorial responsibility"; an on-demand music video service. It expressly notes that a user-to-user service lies outside the regime, as do online newspapers, a private service (eg within a business intranet) and "an individual's personal social media account which is used to share video content with friends and family". Ofcom has held that some channels on YouTube constitute ODPS, though this issue has not been raised that often so it is hard to predict which other channels could be caught.

19. Section 365E(1) Communications Act states:

> "An on-demand programme service must not contain any material likely to incite hatred based on race, sex, religion or nationality".

There is little guidance on what this means specifically as regards ODPS. Ofcom has issued a number of rulings as regards broadcast services (community radio and satellite tv), distinguishing between rulings on content that is discriminatory and that which is hateful. In so far as the broadcast rulings are relevant for understanding the ODPS rules (a point on which there are no decisions), it would seem that the hate rulings would indicate the relevant standard for ODPS. The Broadcasting Code defines "hate speech" as: "all forms of expression which spread, incite, promote or justify hatred based on intolerance on the grounds of disability, ethnicity, gender, gender reassignment, nationality, race, religion, or sexual orientation". These categories of content are important if we understand that part of misinformation operations includes the spread of divisive material.

20. Further, an ODPS must not contain any specially restricted material unless the material is made available in a manner which secures that persons under the age of 18 will not normally see or hear it.

> "Specially restricted material" means—
>
> (a) a video work in respect of which the video works authority has issued a R18 classification certificate;

(b) material whose nature is such that it is reasonable to expect that, if the material were contained in a video work submitted to the video works authority for a classification certificate, the video works authority would issue a R18 classification certificate; or

(c) other material that might seriously impair the physical, mental or moral development of persons under the age of 18.

21. This links the regime back to BBFC standards, which may cover content considerations beyond pornography. Note that the Online Safety Act pornography provisions would not apply where a service provider falls within the ODPS rules.

22. In general, Ofcom seems not to enforce the ODPS rules particularly heavily - perhaps because of early political signals or because of resources. The regime does not seem to be particularly visible to members of the public. If some television channels move purely online, it could be that (if established in the UK) they would fall within the ODPS rules. While hate speech is not acceptable there are currently no rules around accuracy or news, nor impartiality.

The Media Bill

23. Among other objectives, the Media Bill aims to introduce new controls around video on demand (VOD) (part 4 and sch 5-7). It also updates the prominence rules for public service broadcasters. This latter set of rules covers a limited range of devices and does not address prominence in search and social media. This arguably a gap, especially if part of the solution to concerns around mis and disinformation relates to highlighting accurate/authoritative sources of information.

24. Clauses 37 to 40 Media Bill, together with schedules 5 to 7, extend the jurisdiction of the VOD rules to some non-UK based service providers and would provide Ofcom with new regulatory powers to draft and enforce a VOD code applying to them as well as to some, but not all, UK-based ODPS. All UK ODPS seem still to be subject to s 368E- 368H which outline content standards for both editorial and commercial content.

25. The amendments to the ODPS regime which came in as a response to Brexit currently limit Ofcom's jurisdiction to VOD services that

have their head office in the UK; and editorial decisions are taken in the UK.

26. The Media Bill extends jurisdiction to VOD services which do not satisfy these two requirements if:

the members of the public for whose use it is made available are or include members of the public in the United Kingdom.

27. Services which satisfy these requirements each constitute a "non-UK on-demand programme service". Specific obligations apply to "Tier 1 Services". The details of these obligations are found in Sch 5-7 Media Bill.

28. Essentially current sections 368E (harmful material), 368F (advertising), 368FA (HFSS), 368G (sponsorship) and 368H (product placement) Communications Act would apply to a sub-set of non-UK on-demand programme service known as Tier 1 (these are to be designated by the Secretary of State), though the regulations may also specify that some aspects of the rules do not apply to specific services. It is not clear which services will be covered: Netflix, Amazon Prime and Disney+ are mentioned in press releases; Ofcom's roadmap refers to services which are "tv-like" (although this requirement was removed from the Communications Act in 2018 in relation to UK ODPS generally). For non-UK based Tier 1 services, the duties apply "only so far as they are made available for use by members of the public in the United Kingdom". This brings these non-UK services into line with standards currently applicable to UK services.

29. Under the proposed s 368HF Communications Act, (Media Bill p141) Ofcom is required to produce a code to attain "the standards objectives", listed as

> (a) that persons under the age of 18 are protected;
>
> (b) that material likely to encourage or incite the commission of crime or to lead to disorder is not included in Tier 1 services;
>
> (c) that news included in those services is presented with due impartiality;
>
> (d) that news included in those services is reported with due accuracy;
>
> (e) that the impartiality requirements described in section 368HG are met;
>
> (f) that generally accepted standards are applied to the contents of those services so as to provide adequate protection for members of the public from the inclusion of offensive and harmful material;
>
> (g) that the proper degree of responsibility is exercised with respect to the content of religious programmes included in those services (and there are further requirements around exploitation of susceptibilities of audiences here).

30. The new section 368HJ which will be inserted in the Communications Act provides that a Tier 1 service must observe the code. Ofcom is to establish a complaints mechanism in relation to the code. There are additional requirements around accessibility for those with disabilities. The code on privacy (relating to broadcasting) found in s 107 Broadcasting Act 1990 seems to be extended to Tier 1 services.

31. In addition to some non-UK ODPS, Tier 1 includes ODPS services provided as part of a public service remit (with the exception of the BBC which is subject to separate rules). So these additional rules will not apply to all UK based ODPS providers.

Appraisal

32. Our assessment of these developments is that the regime will seem to introduce a three tier system:

- Tier 1 services will be subject to the most detailed rules – rules which look a lot like the broadcasting standards
- Other UK ODPS will be subject to the base level rules found in s 368E- 368H Communications Act
- Non-UK ODPS which are not Tier 1 are outside the scope of regulation even though they are watched in the UK (though obviously, the UK would not want to set standards for the entire globe).

33. The degree of protection in relation to both UK-based and non-UK based services depends to a large degree on the scope of Tier 1 about which there is currently no information. Significantly, news and current affairs rules would apply to them, reducing the risk of distorted and inaccurate reporting. For other services within the UK, hate speech at least would be excluded. However there is still a question as to what approach Ofcom will take to enforcement.

The limitations of the Online Safety Act

34. The Online Safety Act received Royal Assent in October 2023 but its impact - and the powers that Ofcom might be able to swiftly deploy in relation to any emerging election-related risks or threats in the coming months - are likely to be limited in relation to threats to democracy. In that light, we are concerned that in answer to a PQ on the role of the Defending Democracy taskforce on 5 February this year, the Home Office Minister Tom Tugendhat said "The Online Safety Act places new requirements on social media platforms to swiftly remove illegal misinformation and disinformation - including artificial intelligence-generated deepfakes - as soon as they become aware of it." As Full Fact, who campaigned strongly for more robust measures to tackle mis and disinformation in the BIll during its passage through Parliament have observed "There is no credible plan to tackle the harms from online misinformation in the Online Safety Act and this continues to leave the public vulnerable and exposed to online harms. The only references to misinformation in the Act are about setting up a committee to advise Ofcom and changes to Ofcom's media literacy policy. The Act does not address health misinformation, which was so harmful during the Covid-19 pandemic. It also does not set out any new provisions to tackle election disinformation (unless it is a foreign interference offence), nor misinformation that happens during 'information incidents' when information spreads quickly online, such as during terror attacks. The Online Safety Act also does not extend to most harms from the kind of generative AI misinformation we have seen increase in recent months."

35. While we agree in broad terms with this summary, we set out below some more detailed analysis as to how the Act might bite on aspects of the mis/disinformation environment and highlight some of the policy decisions made by the Government during the development and passage of the Online Safety Bill that have led to its fairly narrow application.

36. The Government's initial policy proposals, set out in the 2019 Online Harms White Paper, included disinformation within the scope of harms it sought to address. The Act, however, does not directly target misinformation, disinformation or malinformation as categories of content - illegal or otherwise.

37. The only mention of disinformation/misinformation in the Online Safety Act is at section 152, related to the establishment of a committee to advise Ofcom on the topic. Yet, despite this, it would not be true to say that the OSA does not tackle misinformation or disinformation at all. To understand this, it is necessary to consider how the OSA categorises information and how that intersects with the concepts of disinformation and misinformation. We need also note that the fight against poor quality information is not just about its suppression of but potentially about media literacy (or more transparency about information) as well as the promotion or valuing of good quality information. We have covered some of these aspects in our response to the question above and provide more detail in relation to the existing non-legislative initiatives below.

38. The OSA imposes on providers of user-to-user services and search services duties in relation to two different types of content: illegal content and content harmful to children. For each category of content, there is a risk assessment duty and a 'safety duty' – that is a duty to mitigate the risk of harms arising from content. The duty to mitigate is not just about having systems in place that allow the take down of individual items of content (and note that Ofcom does not have the power to direct that specific items of content be taken down), but includes considering the impact of the service design (e.g. recommender tools and other features affecting the virality of content) and user base. In addition to these core duties, there are provisions applicable to some user-to-user and search services relating to fraudulent content. There are also various other duties relating to transparency and providing reporting and complaints procedures.

39. The regime for each category of content identifies specific examples in respect of which more specific mitigation obligations are imposed: these are the priority illegal offences (listed in Schedules 5-7 OSA) and priority content in relation to children. As regards priority illegal content, the inclusion of the new Foreign Interference Offence (discussed below) brings deliberate activities to subvert democracy by hostile states into scope of the Act. (Ofcom has carried out research into the role of search services in relation to that new offence here.) When the Government conceded that priority content harmful to children should be specified on the face of the legislation and introduced amendments specifying two types of priority content harmful to children: "primary priority content" and "priority content", the reference to health misinformation, found in an earlier Written Ministerial Statement was not included in either list.

40. Ofcom published its first consultation on the regulatory products associated with the illegal content duties in November 2023; the second round of consultation relates to the children's duties and is expected in May this year. The role of the guidance, codes and enforcement notices will be significant in determining the approach within the regime. (We draw the Committee's attention to our response to the consultation in which we raise a number of

concerns about the approach being proposed by Ofcom.) Given there is no express reference to misinformation here, it is unlikely that there will be a separate code on misinformation or disinformation. Presumably, Ofcom's Committee on Disinformation, when established, might provide Ofcom with advice in this area, and this advice could influence other codes of practice. While the Lords, during the passage of the OSB, were keen to see the establishment of this Committee expedited (see debate here), there is as yet no timescale from Ofcom as to when this will happen.

41. As regards content harmful to adults, some user-to-user services known as category 1 (but not search) are required to enforce their terms of service (and cannot go beyond those terms). This could affect misinformation in some instances. While the Government response to the Online Harms White Paper suggested that there would be minimum terms of service, at least surrounding priority content, this requirement does not appear in the OSA. Category 1 services must also provide "user empowerment tools" for adults with regard to types of non-illegal content specified in the OSA, including abuse based on a protected characteristic. Some abuse containing harmful tropes could be seen as a form of disinformation – it would certainly cover hate speech that does not reach the criminal threshold. Relying on user empowerment tools does not however address the societal harms arising from the spread of misinformation – it merely means that that those people who so choose need not see some content. It does not mean that the content is not there, nor does it affect the virality of that content.

Minimising the Spread of Misinformation

Illegal Content Duties

42. The National Security Act 2023- introduced a new criminal offence of foreign interference which was designated a priority offence for the purposes of the OSA. By turning the offence into a priority offence, the OSA imposes an obligation on a service provider to consider foreign interference in its risk assessment, taking into account, in particular, the level of risk of the service being used for the commission or facilitation of the offence, the impact of functionalities of the service and how the design and operation of the service may reduce or increase the risks. Under the safety duty, user-to-user services have an obligation to design a service with the objective of preventing users from encountering this sort of content, and manage the risk of the service being used to facilitate the commission of the offence. Systems should also be aimed at minimising the length of time for which foreign interference material is present – as well as function so as to allow the swift removal of such content when given notice of it.

43. While the OSA does envisage systems to take down illegal content as one possibility, the duties here (as for other forms of priority offence) are not limited to take down. The Government emphasised the importance of tools to control virality. The then Minister for Digital specifically noted "this could include measures to ensure that platform manipulation—such as misleading users about the ownership of an account, or artificially coordinated messaging campaigns—is more difficult, thus mitigating the risk of platform manipulation and disinformation more broadly". The Minister for Security noted "[d]isinformation is often seeded

by multiple fake personas, with the aim of getting real users, unwittingly, then to 'share' it. We need the big online platforms to do more to identify and disrupt this sort of coordinated inauthentic behaviour. That is what this proposed change in the law is about".

44. Ofcom's consultation on illegal content reinforces this point. While there has been a lot of emphasis on recommender systems' roles in spreading illegal content, Ofcom also notes a number of other functionalities on U2U services which can be exploited for foreign interference purposes. This includes: networks of fake accounts or coordinated networks and groups; use of bots; and the ability – especially on messaging services – to forward content. Hyperlinks are used to facilitate cross platform disinformation activities, particularly permanent links. It also notes that Wikipedia's editing processes can be exploited for foreign interference, but features allowing the editing or manipulation (eg deepfakes) of content raise concerns. Ofcom also commented that "services offering advertising, without effective moderation policies to identify and label adverts that seek to influence public political [foreign influence] opinion as 'political', are more able to be used by malicious advertisers". Mitigations dealing with these upstream issues could include measures such as making it more difficult to create large-scale fake accounts or tackling the use of bots in malicious disinformation campaigns. Unfortunately, as set out in our Illegal Harms consultation response, the measures Ofcom is prepared to recommend (which essentially place a ceiling on the service providers' obligations) have not been very far-reaching.

45. Hate speech and harassment are some of the other priority offences listed in Schedule 7 OSA. Insofar as misinformation is included or comprises this sort of behaviour, it will be caught by the regime. Hate speech against women is not yet criminalised, although some forms of harms particularly affect women – for example deepfakes to undermine women in public life. Ofcom has noted, however, that activities covered by the priority offence of foreign interference have targeted those with protected characteristics giving the specific example of foreign interference targeting women in power to foster gendered narratives and expectations that undermine their power. By contrast, user empowerment tools do cover content that contains gender-based hatred.

46. In addition to priority offences, there are general mitigation duties in relation to non-designated illegal content. This is content the use, possession, viewing, accessing, publication or dissemination of which is a criminal offence, and where the victim of the crime is an individual. Given the breadth of coverage, the obligations are less onerous and specific. The central duty is to "effectively mitigate and manage the risks of harm to individuals as identified in the most recent illegal content risk assessment", though this must be understood in the light of any relevant Ofcom codes.

47. The OSA introduces some new criminal offences, including a provision replacing Section 127(2) Communications Act 2003. This provision, section 179 OSA, contains the false communications offence based on the proposals of the Law Commission of England and Wales. The offence covers the knowing sending of inaccurate information where there is an intention to cause "non-trivial psychological or physical harm". Significantly, a court could not find

someone guilty of the false communications offence if that person has a reasonable excuse, for example if the communication was or was intended as a contribution to the public interest, and it is for the prosecution to show that there was no reasonable excuse. A "recognised news publisher" (as defined in the OSA) cannot commit this offence. This seems to focus the offence on disinformation rather than misinformation. The Government position was that this offence will cover behaviours such as deliberately sharing dangerous inaccurate information relating to COVID-19 treatments. While this offence is not designated as a priority offence, it could be a relevant offence for the purposes of the illegal content duties and Ofcom has included it in its consultation on Illegal Content, comparing it to the foreign interference offence. Many of the functionalities noted as risk factors for that offence are similarly relevant. There is a gendered component here, too.

48. We noted in our consultation response that Ofcom has limited the measures recommended in its draft code on illegal content largely to post-hoc, reactive measures such as content moderation and takedown rather than more systemic, upstream interventions. Insofar as these measures will – when the code of practice is in force later in 2024 (too late, most likely, for the UK General Election) – have an impact on the during this period, it is helpful to see that Ofcom has specifically recommended that:

> "Services which are large or multi-risk should resource their content moderation functions so as to give effect to their internal content policies and performance targets, having regard to: the propensity for external events to lead to a significant increase in demand for content moderation on the service; and the particular needs of its United Kingdom user base as identified in its risk assessment, in relation to languages." (Vol 4, 12.45 e)

49. Elsewhere, as noted above, Ofcom includes elections in the category of "sensitive events", which may be a target for foreign interference. So, with respect to offences which meet the criminal threshold relating to hate, intimidation, false communications and foreign interference, an increase in content moderation resources will be required – from large or multi-risk platforms at least – during an election period. The additional measures set out in our response on the question above could be part of this enhanced content moderation, to take into account the specific risks faced by candidates during an election period.

50. More generally, platforms may choose to take action against misinformation, but they must specify what the rules are in their terms of service (which should also be clear). Relying on this as a mechanism to deal with misinformation has some obvious problems. Platforms may simply choose not to deal with some types of misinformation or, even if they have policies against misinformation, those policies may be inadequate – badly defined, for example, or not sufficiently well elaborated. Insofar as hateful content is concerned, there are concerns that the policies are drafted to exclude "comments that are likely to make people leave a discussion" rather than whether they are contributing to stereotypical inaccurate tropes – though presumably these too would make (some) people leave the discussion. Automated mass moderation tools may have difficulty addressing misinformation (including hate speech) within

acceptable tolerances for false positives/false negatives and tend to be orientated towards English (and possibly other major languages e.g. Spanish). Although English is the dominant language in the UK, it is not the only language spoken.

Promotion of Media Literacy

51. The Government via the OSA has expanded Ofcom's media literacy obligations, requiring Ofcom to take steps to heighten the public's awareness and ways in which they can protect themselves. This might also be strengthened by the advice Ofcom receives from the advisory committee on misinformation. Nonetheless, focussing on assessing information that others create overlooks the role of media literacy in what users choose to create. It also overlooks the potentially corrosive effect chronic mistrust might have; potentially reliable sources could be treated with the same disbelief as misinformation. Finally, there is a question as to the context in which people are expected to be media literate; to what extent do providers enable or undermine media literacy through the way they design services, especially for young people who are growing up in a chaotic information environment and obtaining the majority of their news, current affairs and public interest information via social media. Section 11 Communications Act does impose a requirement on Ofcom to "encourage the development and use of technologies and systems for regulating access to [electronic] material", though, as yet, Ofcom seems not to have been effective in deploying this power in relation to misinformation and the new provisions envisage the development of tools for users to protect themselves. Neither of these really tackles the underlying concerns about the business model (e.g. advertising revenue incentives for clickbait content affecting content creators or the need for user engagement affecting platform design).

52. One unusual provision pertaining to media literacy, which could be relevant to misinformation, is found in Section 175 OSA. It empowers the Secretary of State to give Ofcom directions to prioritise content related to specified threats in the exercise of their media literacy powers. The relevant threats relate to the health and safety of the public or to national security, some of which could relate to misinformation. The Explanatory Notes to the Bill gave the example of the Secretary of State issuing during a pandemic a direction requiring Ofcom to give priority to ensuring that health misinformation and disinformation are effectively tackled. This is supported by a further power for the Secretary of State to direct Ofcom to require specified service providers to make a public statement as to identifying the steps they are taking to respond to a specified threat. This power is in addition to the transparency and information obligations supporting the safety duties. While the power is limited to certain situations, these are quite broadly drawn and not limited to emergencies or immediate threats.

Content of democratic importance

53. We note that there is also an obligation in the Act for category 1 services to take proportionate measures to protect content of democratic importance (section 17). As yet, Ofcom has provided no guidance on this and it is unclear to us whether this provision implies a requirement to protect the speakers (eg MPs or other elected officials) from threats that might silence them.

Conclusion

54. We hope that the analysis above has been helpful to the Committee in its deliberations and would be delighted to provide further information or evidence as the inquiry continues.

March 2024

Contact: maeve@onlinesafetyact.net