CarnegieUK TRUST

CHANGING MINDS • CHANGING LIVES

Response To Irish Government Consultation on The Regulation of Harmful Content on Online Platforms and The Implementation of the Revised Audiovisual Media Services Directive

April 2019

Introduction

1. Since early 2018, Professor Lorna Woods and William Perrin have been working, under the aegis of Carnegie UK Trust, on a detailed regulatory proposal for internet harm reduction. As the Irish government considers the development of an Online Safety Act, we hope that this submission on our work provides some useful material for consideration.

2. Carnegie UK Trust was established in 1913 by Scottish-American industrialist and philanthropist Andrew Carnegie to seek:

   "Improvement of the well-being of the masses of the people of Great Britain and Ireland by such means as are embraced within the meaning of the word "charitable" and which the Trustees may from time to time select as best fitted from age to age for securing these purposes, remembering that new needs are constantly arising as the masses advance."

3. Carnegie UK Trust works across the UK and Ireland to influence policy and deliver innovative practice. Our proposal has been developed with reference primarily to the UK regulatory and legislative system so we do not comment in detail in this submission on, for example, the consultation questions relating to the introduction of the Audio-Visual Media Services Directive (AVMSD) within Ireland but we noted within the British context that the statutory duty of care could be used as a mechanism to implement the video-sharing platform provisions of the AVMSD.

4. We have included references to our more detailed work throughout this submission and would be happy to provide further information or discuss with Irish officials, if helpful.

Background

5. Lorna Woods (Professor of Internet Law, Essex University) and William Perrin (Trustee of Carnegie UK Trust) have been working with Carnegie UK Trust (CUKT) to design a regulatory system to reduce harm on social media. The proposals have been published via a series of blog posts [1] and a new paper has recently been published which consolidates our thinking, and updates some of the work in the light of feedback and discussions with diverse stakeholders over the past 18 months.[2]

1 https://www.carnegieuktrust.org.uk/project/harm-reduction-in-social-media/

2 https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf

6. We have vast experience in regulation, privacy and free speech issues. William has worked on technology policy since the 1990s, was a driving force behind the creation of OFCOM and worked on regulatory regimes in many economic and social sectors while working in the UK government's Cabinet Office. He ran a tech start up and is now a trustee of several charities. Lorna is Professor of Internet Law at University of Essex, an EU national expert on regulation in the TMT sector, and was a solicitor in private practice specialising in telecoms, media and technology law.

7. Our Carnegie work was catalysed by the harms set out in the UK government's Green Paper[3] and much reporting of harms by interest groups. We published our initial work just before the government's May 2018 announcement that they would bring forward a White Paper to:

   'set out plans for upcoming legislation that will cover the full range of online harms, including both harmful and illegal content. Potential areas where the Government will legislate include the social media code of practice, transparency reporting and online advertising.'[4]

8. Our proposals put forward a new regulatory regime to reduce harm to people from social media services. We draw on the experience of regulating in the public interest for externalities and harms to the public in sectors as diverse as financial services and health and safety. Our approach – looking at the design of the service – is systemic rather than content-based, preventative rather than palliative.

9. At the heart of the new regime would be a 'duty of care' set out by Parliament in statute. This statutory duty of care would require most companies that provide social media or online messaging services used in the UK to protect people in the UK from reasonably foreseeable harms that might arise from use of those services. This approach is risk-based and outcomes-focused. A regulator would have sufficient powers to ensure that companies delivered on their statutory duty of care. Our proposals have been adopted as the basis of recommendations to Government by many campaigning organisations and Parliamentary committees in recent months.[5]

10. The long-delayed Online Harms White Paper was published by the UK Government on 8th April 2019 and proposes the introduction of a statutory duty of care, enforced by an independent regulator, which draws on our thinking. We are currently reviewing the detail of their proposals and will respond formally as part of the three-month consultation.[6]

11. We note that the Irish government has committed to bringing forward legislation with similar objectives to those set out by the UK government:

---

3 https://www.gov.uk/government/consultations/internet-safety-strategy-green-paper

4 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/708873/Government_Response_to_the_Internet_Safety_Strategy_Green_Paper_-_Final.pdf

5 A duty of care has been recommended by the following organisations to date in 2019: NSPCC (Taming the Wild West Web, February 2019: https://www.nspcc.org.uk/globalassets/documents/news/taming-the-wild-west-web-regulate-social-networks.pdf); Children's Commissioner: https://www.childrenscommissioner.gov.uk/2019/02/06/childrens-commissioner-publishes-a-statutory-duty-of-care-for-online-service-providers/; UK Chief Medical Officers: https://www.gov.uk/government/publications/uk-cmo-commentary-on-screen-time-and-social-media-map-of-reviews); House of Commons Science and Technology Committee: https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/822/82202.html; UK Labour Party: https://labour.org.uk/press/tom-watson-speech-fixing-distorted-digital-market/; Lords Communications Committee: https://www.parliament.uk/business/committees/committees-a-z/lords-select/communications-committee/inquiries/parliament-2017/the-internet-to-regulate-or-not-to-regulate/; All-Party Parliamentary Group on Social Media and Young People's Mental Health and Wellbeing/Royal Society of Public Health: https://www.rsph.org.uk/our-work/policy/wellbeing/new-filters.html)

6 https://www.gov.uk/government/consultations/online-harms-white-paper

'The situation at present where online and social media companies are not subject to any oversight or regulation by the state for the content which is shared on their platforms is no longer sustainable. I believe that the era of self-regulation in this area is over and a new Online Safety Act is necessary'.[7]

12. Our proposals for a statutory duty of care are therefore as relevant to the consideration of the scope and objectives of online legislation in Ireland as they are in the UK. Indeed, we briefed Donnchadh Ó Laoghaire TD on our approach last autumn. We also note that, last autumn, Senator Catherine Noone called for a "statutory duty of care" for social media companies to address online harassment and abuse.[8]

13. The content of this submission is most relevant to the proposals and questions set out by the Department of Communications, Climate Action and Environment in Strand 1 of the consultation's Explanatory Note (new online safety laws to apply to Irish residents). Please see our more detailed work for consideration of the intersection with the Audio Visual Media Services Directive (AVMSD) and other European legislation, which will also have relevance to the Irish regulatory context.[9]

A statutory duty of care for online harm reduction

14. We note that the consultation starts from the premise of designing an appropriate system "*to require the removal of harmful content from online platforms*" and ask whether this requires the "*regulatory oversight of the notice and take down systems which services have in place, or the direct involvement of the regulator in a notice and take down system where it would have a role in deciding whether individual pieces of content should or should not be removed?*"[10]

15. We would argue that designing a regulatory system around take-down systems for content is too reactive; it will also be problematic in relation to the risk that the consultation rightly flags "that legitimate freedom of speech and freedom of expression online could be curtailed."

16. Our proposal draws on the long-established precautionary principle for policymaking, coupled with a "safety by design" approach, which put the onus instead on social media platforms to prevent users from the risk of coming to reasonably foreseeable harm on their platforms. Takedown procedures are a necessary part of the activity overseen and monitored by the regulator but by starting from an outcome-focused, risk-based position, we believe that greater protection can be afforded to users, particularly vulnerable groups.

7 'Minister Bruton proposes new law to protect children online'; Press release: 5 March 2019

8 https://www.irishexaminer.com/breakingnews/ireland/catherine-noone-wants-statutory-duty-of-care-for-social-media-firms-to-fight-online-abuse-870903.html

9 See our recent paper: (https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf), chapter 4.

10 Department of Communications, Climate Action and Environment: "Public Consultation on the Regulation of Harmful Content on Online Platforms and the Implementation of the Revised Audiovisual Media Services Directive: Explanatory Note" March 2019 p4-5

Platforms as public spaces

17. Social media and messaging service providers should each be seen as responsible for a public space they have created, much as property owners or operators are in the physical world. Everything that happens on a social media or messaging service is a result of corporate decisions: about the terms of service, the software deployed and the resources put into enforcing the terms of service. These design choices are not neutral: they may encourage or discourage certain behaviours by the users of the service.

18. In the physical world, the UK Parliament has long imposed statutory duties of care upon property owners or occupiers in respect of people using their places, as well as on employers in respect of their employees. A statutory duty of care is simple, broadly based and largely future-proof. For instance, the duties of care in the Health and Safety at Work Act 1974[11] work well today, enforced and with their application kept up to date by a competent regulator. A statutory duty of care focuses on the objective – harm reduction – and leaves the detail of the means to those best placed to come up with solutions in context: the companies who are subject to the duty of care. This is the fundamental difference between our approach and the starting point of the questions on notice and takedown that are posed in consultation on the Irish Online Safety Act.

19. A statutory duty of care returns the cost of harms to those responsible for them, an application of the micro-economically efficient 'polluter pays'[12] principle. The E-Commerce Directive[13], permits duties of care introduced by Member States; the Audiovisual Media Services Directive (as amended in 2018) requires Member States to take some form of regulatory action in relation to a sub-set of social media platforms – video-sharing platforms – and we note that the consultation is looking for detailed responses on how this might be implemented in Ireland.

The precautionary principle

20. Rapidly-propagating social media and messaging services, subject to waves of fashion amongst young people in particular, are an especial challenge for legislators and regulators. The harms are multiple, and may be context- or platform- specific, while the speed of their proliferation makes it difficult for policymakers to amass the usual standard of long-term objective evidence to support the case for regulatory interventions. The software that drives social media and messaging services is updated frequently, often more than once a day. Facebook for instance runs a 'quasi-continuous [software] release cycle'[14] to its webservers. The vast majority of changes are invisible to most users. Tweaks to the software that companies use to decide which content to present to users may not be discernible. Features visible to users change regularly. External researchers cannot access sufficient information about the user experience on a service to perform long-term research on service use and harm. Evidencing harm in this environment is challenging, traditional long-term randomised control trials to observe the effect of aspects of the service on users or others are nearly impossible without deep co-operation from a service provider.

11 https://www.legislation.gov.uk/ukpga/1974/37

12 Polluter Pays – OECD 1972 see https://www.oecd-ilibrary.org/environment/the-polluter-pays-prinIciple_9789264044845-en

13 Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L 178/1, available: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32000L0031.

14 Rapid release at massive scale August 2017 https://code.fb.com/web/rapid-release-at-massive-scale/

21. Nonetheless there is substantial indicative evidence of harm both from advocacy groups and more disinterested parties. In the UK, OFCOM (the communications sector regulator) and the Information Commissioner's Office (ICO) have demonstrated[15] that a basic survey approach can give high level indications of harm as understood by users. So, how do regulation and economic activity proceed in the face of indicative harm but where scientific certainty cannot be achieved in the time frame available for decision making?

22. Being called on to act robustly on possible threats to public health before scientific certainty has been reached is not new for Governments. After the many public health and science controversies of the 1990s, the UK government's Interdepartmental Liaison Group on Risk Assessment (ILGRA) published a fully worked-up version of the precautionary principle for UK decision makers.[16]

    'The precautionary principle should be applied when, on the basis of the best scientific advice available in the time-frame for decision-making: there is good reason to believe that harmful effects may occur to human, animal or plant health, or to the environment; and the level of scientific uncertainty about the consequences or likelihoods is such that risk cannot be assessed with sufficient confidence to inform decision-making.'

23. The ILGRA document advises regulators on how to act when early evidence of harm to the public is apparent, but before unequivocal scientific advice has had time to emerge, with a particular focus on novel harms. The ILGRA's work is still current and hosted by the UK Health and Safety Executive (HSE), underpinning risk-based regulation of the sort we propose.

System not content

24. In our most recent paper, we set out in detail our view that online environments reflect choices made by the people who create and manage them; those who make choices should be responsible for the reasonable foreseeable risks of those choices. Our approach – moving beyond a focus on platform accountability for every item of content on their respective sites, to their responsibility for the systems that they have designed for public use – allows the reduction of harm to be considered within a regulatory approach that scales but, by not focussing directly on types of content, is also committed to the preservation of free speech.

25. In the regime we propose, oversight would be at a system or platform level, not regulation of specific content – this is significant in terms of placing responsibility on the actions and activities that platform operators control and also in terms of practicality. Regulation at the system level focuses on the architecture of the platform. This is similar to the 'by design' approach seen in data protection and information security (for example in the EU GDPR). Ongoing review of this design is important to ensure that the system continues to function as the market and technology develops[17]. The statutory duty of care approach is not a one-off action but an ongoing, flexible and future-proofed responsibility that can be applied effectively to fast-moving technologies and rapidly emerging new services.

15 https://www.ofcom.org.uk/__data/assets/pdf_file/0018/120852/Internet-harm-research-2018-report.pdf

16 http://www.hse.gov.uk/aboutus/meetings/committees/ilgra/pppa.htm

17 On difficulties with 'baked in' assumptions see e.g. Antignac, Th, and Le Métayer D., 'Privacy by design: From technologies to architectures' in Preneel B. and Ikonomou D. (eds.) Privacy Technologies and Policy, Springer International Publishing, 2014, pp. 1-17; Koops, B-J., and Leenes, R., 'Privacy Regulation cannot be hardcoded. A critical comment on the 'privacy by design' provision in data protection law' (2014) 28 International Review of Law, Computers and Technology 159.

26. In broad terms, our system-level approach looks like this: the regulator would have powers to inspect and survey the networks to ensure that the platform operators had adequate, enforced policies in place. The regulator, in consultation with industry, civil society and network users would set out a model process for identifying and measuring harms in a transparent, consultative way. The regulator would then work with the largest companies to ensure that they had measured harm effectively and published harm reduction strategies. Specifying the high-level objectives to safeguard the general public allows room for service providers to act by taking account of the type of service they offer, the risks it poses (particularly to vulnerable users) and the tools and technologies available at the time. The approach builds on the knowledge base of the sector and allows for future proofing. The steps taken are not prescribed but can change depending on their effectiveness and on developments in technologies and their uses.

Defining harms

27. The consultation on the proposals for the Online Safety Act asks two important questions on:

- the content that should be considered "harmful content" and the nature of services that should be in scope; and
- the list proposed by the Irish government for defining "harmful content".

28. We have considered similar questions in relation to the introduction of a statutory duty of care and set out our current thinking in summary here. There is more detail on both areas in our recent paper.

29. We note that the consultation proposes the following example of the types of material that could be included in a definition of "Harmful content": serious cyber bullying of a child, material which promotes self-harm or suicide, and material designed to encourage prolonged nutritional deprivation that would have the effect of exposing a person to risk of death or endangering health. We also note the important clarification that online platforms are already required to remove content which is a criminal offence under Irish and EU law.

30. Under our duty of care proposal, we set out a role for Parliament to guide the regulator with a non-exclusive list of harms for it to focus upon. These would include:

- Harmful threats – statement of an intention to cause pain, injury, damage or other hostile action such as intimidation. Psychological harassment, threats of a sexual nature, threats to kill, racial or religious threats known as hate crime. Hostility or prejudice based on a person's race, religion, sexual orientation, disability or transgender identity. We would extend hate crime to include misogyny currently being reviewed by the UK government[18].
- Economic harm – financial misconduct, intellectual property abuse, passing off and consumer scams
- Harms to national security – violent extremism, terrorism, state sponsored cyber warfare

18 Review announced following campaign by Stella Creasy MP https://www.bbc.co.uk/news/uk-politics-45423789

- Emotional harm – such harm is unlikely to be covered by the law as it stands. The common law sets a very high bar for emotional harm, of a 'recognised psychiatric injury'[19]. This is out of kilter with the past decade of the law changing to recognise an emotional component of crime, often where women are victims – stalking, domestic abuse, harassment[20], controlling or coercive behaviour[21]. We suggest that emotional harm is reasonably foreseeable on some social media and that services should have systems in place to prevent emotional harm suffered by users such that it does not build up to the criminal threshold of a recognised psychiatric injury. For instance, through aggregated abuse of one person by many others in a way that would not happen in the physical world or service design that is intentionally addictive. This includes harm to vulnerable people – in respect of suicide, self-harm anorexia, mental illness etc.

- Harm to young people – bullying, aggression, hate, sexual harassment and communications, exposure to harmful or disturbing content, grooming, child abuse (See UKCCIS[22] Literature Review).

- Harms to justice and democracy – prevent intimidation of people taking part in the political process beyond robust debate, protecting the criminal and trial process.

Who should be regulated?

31. Our thinking has evolved here during the course of our work, and we have broadened out our scope beyond an initial definition of social media services. We propose regulating services that:

    - Have a strong two-way or multiway communications component;

    - Display user-generated content publicly or to a large member/user audience or group.

    - Are not subject to a detailed existing content regulatory regime, such as the traditional media.

32. As well as the large platforms such as Facebook, Twitter, YouTube, LinkedIn, TikTok, Kik etc, we would also include messaging services that have evolved and permit large group sizes and semi or wholly public groups, as well as gaming services that include a social network/messaging function. We have yet to come to a definitive view on search engines. The regime would cover reasonably foreseeable harm that occurs to people who are users of a service and reasonably foreseeable harm to people who are not users of a service.

33. If a service provider targets or is used by a vulnerable group of users (e.g. children), the risk of harm is greater and the service provider should have more safeguard mechanisms in place than a service which is, for example, primarily used by adults and which may have community rules agreed by the users themselves.

19 See Stannard – Sticks and Stones https://pure.qub.ac.uk/portal/en/publications/sticks-stones-and-words-emotional-harm-and-the-english-criminal-law(677b44b4-b703-4e73-832f-466b75d58c42)/export.html

20 The CPS guidance provides a good overview of harassment issues https://www.cps.gov.uk/legal-guidance/stalking-and-harassment

21 Also: https://www.cps.gov.uk/legal-guidance/controlling-or-coercive-behaviour-intimate-or-family-relationship

22 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650933/Literature_Review_Final_October_2017.pdf

Regulatory approach

34. It is not appropriate for us to put forward any views on a suitable regulatory structure for the Irish regime; the audience for our proposal is the UK government and as such we recommend establishing a regulator that is independent of government and that OFCOM is the most appropriate regulatory body in the UK to take on oversight for a new statutory duty of care. We would, however, urge caution in setting up a wholly new body to oversee online safety and harm reduction. The time and resources required to achieve this will significantly delay the delivery of urgent action to reduce harms to vulnerable groups. Giving the regulatory responsibilities – at least in the first instance – to an existing authority can mitigate that risk.

35. In relation to the consultation questions on the powers that could be assigned to the regulator and the sanctions it could impose, we set out below and in the annex our thoughts in relation to a statutory duty of care. As we set out above, a risk-based regime that focuses on the outcome means that the service providers are free to choose how to discharge their responsibilities to reduce harm, and we envisage that the designated regulator would be given substantial freedom in its approach to remain relevant and flexible over time.

36. We suggest the regulator employ a harm reduction method similar to that used for reducing pollution: agree tests for harm, run the tests, the company responsible for harm invests to reduce the tested level, test again to see if investment has worked and repeat if necessary. If the level of harm does not fall or if a company does not co-operate then the regulator will have sanctions. We set out more detail on how the regulator would work in the annex.

37. In a model process, the regulator would work with civil society, users, victims and the companies to determine the tests and discuss both companies harm reduction plans and their outcomes. The regulator would have a range of powers to obtain information from regulated companies as well as having its own research function. The regulator is there to tackle systemic issues in companies and, in this proposal, individuals would not have a right to complain to the regulator or a right of action to the courts for breach of the statutory duty of care. We suggest that a form of 'super-complaint' mechanism, which empowers designated bodies to raise a complaint that there has been a breach of statutory duty, be introduced.

38. We have put some detailed thought into appropriate sanctions and penalties in recent months. We believe that a regulator needs effective powers to make companies change behaviour. Our proposal suggests large fines set as a proportion of turnover, along the lines of the General Data Protection Regulation (GDPR)[23] and Competition Act[24] regimes. We have also made suggestions in our most recent paper of powers that bite on directors personally, such as fines.

The need to act

39. We are of the view that urgent action is needed, especially in a fast-changing sector, creating a tension with a traditional deliberative process of forming legislation and then regulation. We are

23 Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1, available: https://eur-lex.europa.eu/legal-content/EN/TX-T/?qid=1552405263176&uri=CELEX:32016R0679

24 https://www.legislation.gov.uk/ukpga/1998/41/contents

heartened that the UK Government has put forward a statutory duty of care as the centrepiece of their regulatory proposals in the recent Online Harms White Paper but need to further consider the detail when drawing up our consultation response.

40. We are under no illusion that the issues involved in this area are complex: social media companies have a duty to shareholder interests; individuals do not bear the costs of their actions when they use social media leading to strong externalities; rights to freedom of speech are highly valued; the issues cut across different global approaches to regulation, making jurisdiction sometimes unclear; and few governments have seemed willing to take a strong lead. So, we therefore welcome the commitment that the Irish government has made to tackling online harms and hope that, wherever both governments end up with their final proposals, there will be an opportunity to work collaboratively on the implementation and learnings from the approaches. We would be happy to discuss our proposals in detail with Irish officials as they consider the next steps in the policymaking process.

41. In the meantime, we urge both governments to find a route to act quickly and bring forward their legislative proposals as early as is practicably possible. If we wait three or four years the harms may be out of control, which will not be good for society, nor the companies concerned.

Professor Lorna Woods
William Perrin
Maeve Walsh

Contact: maeve@carnegieuk.org

ANNEX: The regulatory process and the role of the regulator

42. Given the huge power of most social media service companies relative to an individual we would appoint a regulator to enforce the duty of care; expecting individuals to assert rights through court processes as a mechanism to control the problems on these platforms is unlikely to work given the costs and stresses of litigation, the asymmetry in knowledge and power between the platforms and individual litigants. The regulator would ensure that companies have measurable, transparent, effective systems in place to reduce harm; to do so it would be provided with information-gathering powers. The regulator would have powers of sanction if companies did not comply with their duty of care.

43. Under our proposals, the regulator would be independent. The regulator would not get involved in individual items of speech and would be bound by the Human Rights Act. In exercising its powers, it must have regard to fundamental human rights and the need to maintain an appropriate balance between conflicting rights. The regulator must not be a censor. Regulators in the UK such as the BBFC, the ASA and OFCOM (and its predecessors) have demonstrated for decades that it is possible to combine quantitative and qualitative analysis of media, neutral of political influence, for regulatory process.

44. Central to the duty of care is the idea of risk. We are not proposing that a uniform set of rules apply across very different services and user bases but that the risks and appropriate responses to those risks are assessed in the light of these differences. Harmful behaviours and risk have to be seen in the context of the platform. In assessing whether a statutory duty of care had been met, the regulator would examine whether a social media service operator has had particular regard to its audience. For example, a mass membership, general purpose service open to children and adults should manage risk by setting a very low tolerance for harmful behaviour, in the same way that some public spaces, such as a family theme park, take into account that they should be a reasonably safe space for all.

45. The risk profile would be different for a specialist site targeted at a more limited audience. Specialist audiences/user-bases of social media services may have online behavioural norms that on a family-friendly service could cause harm but in the community where they originate are not harmful. Examples might include sports-team fan services or sexuality-based communities. This can be seen particularly well with Reddit: its user base with diverse interests self-organises into separate subreddits, each with its own behavioural culture and approach to moderation. Mastodon also has distinct communities each of which sets its own community rules (as opposed to ToS imposed by the provider) within the overarching Mastodon framework. In some sites, robust or even aggressive communications (within the law) could be allowed.

46. One possible benefit of this approach could be that there might be more differentiation between providers and consequently possibly more choice, though we note the strong network effects in this sector. Differentiation between high and low risk services is common in other regulatory regimes, such as for data in the GDPR and is central to health and safety regulation. In those regimes, high risk services would be subject to closer oversight and tighter rules, as we intend here.

Harm reduction cycle

47. We envisage an ongoing evidence-based process of harm reduction where the regulator works with the industry and civil society to create a cycle that is transparent, proportionate, measurable and risk-based. The regulator would prioritise high-risk services, and would only have minimal engagement with low-risk services. We describe a cycle here that relies on consultation and feedback loops with the regulator and civil society.

48. A harm reduction cycle begins with measurement of harms. Here we emphasise that as the companies' performance is to be managed at system level, we do not intend that the effect of social media use on each individual should be measured. Rather what is measured is the incidence of artefacts that – according to the code drawn up by the regulator – are deemed as likely to be harmful (to a particular audience) or if novel could reasonably have been be foreseen to cause harm. We use 'artefact' as a catch all term for types of content, aspects of the system (e.g. the way the recommender algorithm works) and any other factors. We discuss foreseeability below.

49. The regulator would draw up a template for measuring harms, covering scope, quantity and impact. The regulator would then consult publicly on this template, specifically including the qualifying social media services. The qualifying social media services would then run a measurement of harm based on that template, making reasonable adjustments to adapt it to the circumstances of each service. The regulator would have powers in law to require the companies providing the qualifying services to comply. The companies would be required to publish the survey results in a timely manner. This would establish a first baseline of harm.

50. The companies would then be required to act to reduce these harms by setting out and implementing a harm reduction action plan. We expect those planned actions to be in two groups – things companies just do or stop doing, immediately; and actions that would take more time (for instance new code or terms and conditions changes). Companies should inform the regulator and publish their actions. Companies should seek views on their action plan from users, the victims of harms, the NGOs that speak for users and victims etc. The companies' responses to public comment (though they need not adopt every such suggestion made) would form part of any assessment by the regulator of whether an operator was taking reasonable steps and satisfying its duty of care. Companies would be required to publish their action plans, in a format set out by the regulator, such as:

    - what actions they have taken immediately;
    - actions they plan to take;
    - an estimated timescale for measurable effect; and
    - basic forecasts for the impact on the harms revealed in the baseline survey and any others they have identified.

51. The regulator would take views on the plan from the public, industry, consumers/users and civil society and makes comments on the plan to the company, including comments as to whether the plan was sufficient and/or appropriate. The companies would then continue or begin their harm reduction work.

52. Harms would be measured again after a sufficient time has passed for harm reduction measures to have taken effect, repeating the initial process. This establishes the first progress baseline. The baseline will reveal four possible outcomes – that harms:

- have risen;
- stayed the same;
- have fallen; or
- new harms have occurred/been revealed.

53. If harms surveyed in the baseline have risen or stayed the same the companies concerned will be required to act and plan again, taking due account of the views of victims, NGOS and the regulator. In these instances, the regulator may take the view that the duty of care is not being satisfied and, ultimately, may take enforcement action (see below). If incidence of harms has fallen then companies will reinforce this positive downward trajectory in a new plan. Companies would prepare second harm reduction reports/plans as in the previous round but including learning from the first wave of actions, successful and unsuccessful. Companies would then implement the plans. The regulator would set an interval before the next wave of evaluation and reporting.

54. Well-run social media services would quickly settle down to a much lower level of harm and shift to less risky service designs. This cycle of harm measurement and reduction would continue to be repeated, as in any risk management process participants would have to maintain constant vigilance of their systems.

55. We anticipate that well-run services and responsible companies will want to comply with a harm reduction process. Where companies do not comply, or where the regulator has grounds to believe that they have not we propose that the regulator has information gathering powers as is normal in modern regulation (see for example the powers granted to the Information Commissioner in the UK). A net output of the harm reduction cycle would be a transparency report produced by each company to ensure an accurate picture of harm reduction was available to the regulator and civic society organisations.

Measurement

56. While the harm reduction cycle envisages that the use of the platforms by users should be subject to some sort of surveying, we do not envisage that the measurement necessitates constant monitoring of all use and neither statute nor the regulator should require this. At the scale at which many platforms operate, statistical sampling methods should be sufficiently robust combined with other measures. For instance, the platforms' own mailbox/complaints log should provide early warning of systemic problems.

Foreseeability

57. The regulator will need to consider with industry, victims and civil society how foreseeability of material risks applies in the context of social media and messaging. This is at the heart of any risk assessment – there will be risks which will be obvious – for instance material harm is known to have occurred before and those which, while not obvious are foreseeable. If a material risk is foreseeable

then a company should take reasonable steps to prevent it. Health and safety law for instance has developed substantial practice around foreseeability of risk and the regulator in this case will have to take a similar approach.

From Harms to Codes of Practice

58. An output of the harm reduction cycle would be industry generated codes of practice that could be endorsed by the regulator. In our view, the speed with which the industry moves would mitigate against traditional statutory codes of practice which require lengthy consultation cycles. The government, in setting up such a regime, should allow some lee-way from standard formalised consultation and response processes. Codes of practice, as well as other forms of guidance (which could also be produced by the regulator) make compliance easier for small companies.

Proportionality

59. Some commentators have suggested that applying a duty of care to all providers might discourage innovation and reinforce the dominance of existing market players. We do not think that the application of the duty of care would give rise to a significant risk in this regard, for the following reasons.

60. Good regulators do take account of company size and regulation is applied proportionate to business size or capability. We would expect this to be a factor in determining what measures a company could reasonably have been expected to have taken in mitigating a harm. Clearly, what is reasonable for a large established company would be different for an SME. The proportionality assessment proposed does not just take into account size, but also the nature and severity of the harm, as well as the likelihood of it arising. For small start-ups, it would be reasonable for them to focus on obvious high risks, whereas more established companies with greater resources might be expected not only to do more in relation to those risks but to tackle a greater range of harms.

61. The regulator should determine, with industry and civil society, what is a reasonable way for an SME service provider to manage risk. Their deliberations might include the balance between managing foreseeable risk and fostering innovation (where we believe the former need not stymie the latter) and ensuring that new trends or emerging harms identified on one platform are taken account of by other companies in a timely fashion. The regulatory emphasis would be on what is a reasonable response to risk, taken at a general level. In this, formal risk assessments constitute part of the harm reduction cycle; the appropriateness of responses should be measured by the regulator against this.

62. We note that, in other sectors (notably HSWA, data protection and guidance on the Content Codes for broadcasting), regulators give guidance on what is required by the regulatory regime and ways to achieve that standard. This saves businesses the cost of working out how to comply. As in other sectors, regulation will create or bolster a market for training and professional development in aspects of compliance. We would expect the regulator to emphasise the need for training for start-ups and SME`s on responsibility for a company's actions, respect for others, risk management etc. The work on ethics in technology could usefully influence this type of training.

63. Furthermore, regulators would not be likely to apply severe sanctions in the case of a start-up, at least initially. A small company that refused to engage with the regulatory process or demonstrated cavalier behaviour leading to harms would become subject to more severe sanctions. Sanctions are discussed below.

Techniques for harm reduction

64. We do not envisage that a system that relied purely on user notification of problematic content or behaviour and after the event responses would be taking sufficient steps to reduce harm. In line with the 'by design' approach, consideration for the effects of services and the way they are used should happen from the beginning of the design process and not be mere afterthoughts and consider the way the system as designed affects its users and their behaviour. Nor should the responsibility for safety principally fall to individuals affected (some of whom may not even use the platform) to configure complex tools and to be resilient enough to complain (and persist in that complaint in the fact of corporate disinterest), especially where those people are vulnerable. We draw the following from a wide range of regulatory practice, but the list is not intended to be exhaustive. Some of these the regulator would do, others the regulator would require the companies to do if they were not implementing systems to drive down risk of harm.

65. Each qualifying social media service provider could be required to:

- develop a statement of assessed risks of harm, prominently displayed to all users when the regime is introduced and thereafter to new users; and when launching new services or features;
- provide its child protection and parental control approach, including age verification, for the regulator's approval;
- develop easy to use tools for users to control the content they see and to limit their exposure to others;
- display a rating of harm agreed with the regulator on the most prominent screen seen by users;
- an internal review system for risk assessment of new services, new tools or significant revision of services prior to their deployment (so that the risk is addressed prior to launch or very risky services do not get launched);
- develop a triage process for emergent problems (the detail of the problem may be unknown, but it is fairly certain that new problems will be arising);
- work with the regulator and civil society on model standards of care in high risk areas such as suicide, self-harm, anorexia, hate crime etc;
- provide adequate complaints handling systems with independently assessed customer satisfaction targets and also produce a twice yearly report on the breakdown of complaints (subject, satisfaction, numbers, handled by humans, handled in automated method etc.) to a standard set by the regulator; and

- assess how effective the company's enforcement of its own terms of service is and if necessary improve its performance.

66. The regulator would at a minimum use the following tools and techniques:

- publish model policies on user sanctions for harmful behaviour, sharing research from the companies and independent research;
- detailed guidance as to the meaning of harm;
- publish transparency report models for companies to follow;
- set standards for and monitoring response time to queries;
- co-ordinate with the qualifying companies on training and awareness for the companies' staff on harms;
- approve industry codes of practice, where appropriate; and
- provide guidance to companies, in particular with SMEs in mind on suggested approaches to dealing with well-known problems and watchlists of issues that might arise when operating particular types of service (or encourage trade bodies to do so);
- monitor if regulated problems move elsewhere and to spread good practice on harm reduction;
- publish a forward-look at non-qualifying social media services brought to the regulator's attention that might qualify in future;
- support research into online harms – both funding its own research and co-ordinating work of others;
- establish a reference/advisory panel to provide external advice to the regulator – the panel might comprise civil society groups, people who have been victims of harm, free speech groups.

Sanctions

67. Some of the qualifying social media services will be amongst the world's biggest companies. In our view the companies will want to take part in an effective harm reduction regime and comply with the law. The GDPR penalties and sanctions regime (including levelling fines as a proportion of revenue for data breaches, along with the impact of consequent publicity and reputational damage) have yet to be fully exercised by the ICO in the UK and may yet provide an effective preventative model. The impact of the CNIL decision against Google in France will be an early indicator of the effectiveness of the GDPR regime in modifying corporate behaviour.[25]

25 CNIL announcement, 21 January 2019: https://www.cnil.fr/en/cnils-restricted-committeeimposes-financial-penalty-50-million-euros-against-google-llc

68. The companies' duty is to their shareholders. There is an argument that, in order to spend significant shareholder resources on matters for the public good a company management requires regulation. The scale at which these companies operate means that a proportionate sanctions regime is required.

69. The range of mechanisms available within the Health and Safety regime allow the regulator to try improve conditions rather than just punish the operator, so we would propose a similar range of notices (and to some extent the ICO has a similar approach). For those that will not comply, the regulator should be empowered to impose fines (perhaps GDPR magnitude fines if necessary). We have noted in the context of the ICO that a range of investigative powers to support effective enforcement were introduced; we propose that similar powers be given to the regulator here - a comprehensive suite of information gathering powers such as the ICO's ability to make information notices under s. 142 of the Data Protection Act 2018 and information orders under s. 145.

70. All regulatory processes leading to the imposition of sanctions should be transparent and subject to a civil standard of proof. The regulator, like any public body, would be subject to judicial review.

71. Sanctions would include:

- Administrative fines in line with the parameters established through the Data Protection Act/ GDPR regime of up to €20 million, or 4% annual global turnover – whichever is higher. Many types of fines, however, are routinely insured against.

- Enforcement notices – (as used in data protection, health and safety) – in extreme circumstances a notice to a company to stop it doing something. Breach of an enforcement service could lead to substantial fines.

- Enforceable undertakings where the companies agree to do something to reduce harm.

- Adverse publicity orders – the company is required to display a message on its screen most visible to all users detailing its offence. A study on the impact of reputational damage for financial services companies that commit offences in the UK found it to be nine times the impact of the fine.

- Forms of restorative justice – where victims sit down with company directors and tell their stories face to face.

Who should regulate?

72. When considering this, the first question is whether a regulator is needed at all if a duty of care is to be created. Our view is that a regulator can address asymmetries of power between the victim and the harm causer. It is conceivable for a home owner to sue a builder or a person for harm from a building, or a person to sue a local authority for harm at a playground. However, there is a strong power imbalance between an employee and their boss or even between a trade union and a multinational. A fully functioning regulator compensates for these asymmetries. In our opinion,

there are profound asymmetries between a user of a social media service and the company that runs it, even where the user is a business, and so a regulator is required to compensate for the users' relative weakness.

What Sort of Regulator?

73. Assuming a regulator is needed, should it be a new regulator from the ground up or an existing regulator upon which the powers and resources are conferred? Need it be a traditional regulator, or would a self or co-regulator suffice? There are many instances of co-regulation in the communications sector that have run into problems. Self-regulation works best when the public interest to be served and those of the industry coincide. This is not the case here.

74. Our view is that, whichever model is adopted, the important point is that the regulator be independent. The regulator must be independent not only from government but also from industry, so that it can make decisions based on objective evidence (and not under pressure from other interests) and be viewed as a credible regulator by the public. This is particularly important given the fundamental human rights that are in issue in the context of social media. Independence means that the regulator must have sufficient resources, as well as relevant expertise.

75. A completely new regulator created by statute would take some years before it was operational. In our view harm reduction requires more urgent action and for this reason we reject the idea, seen in some proposals, that a new sector specific regulator is required.

76. Instead, our work proposes extending the competence of an existing regulator: it spreads the regulator's overheads further, draws upon existing expertise within the regulator (both in terms of process and substantive knowledge) and allows a faster start. In a UK context, we recommend that OFCOM takes on the powers to reduce harm in social media services: it has long experience in digital issues, a strong research capability and proven independence as well as resilience in dealing with multinational services.