Introducing the Systems Approach and the Statutory Duty of Care

Lorna Woods

Abstract: Early policy in relation to the internet framed questions from the perspective of liability for individual items of content. With the growth of social media, the approach struggles to deal with the scale of material as well as the contextual subjectivity of the acceptability of some types of content. This chapter explains a different approach, based on the work of Carnegie UK Trust, that moves away from direct content regulation to look at the services on which that content is created and disseminated. It argues that those services are not neutral as to that content, and that design choices can operate to create or exacerbate problems. The proposal is that of a risk managed approach to service development, aiming to achieve 'safety by design'. Although the orginal Carnegie proposal was based in English law, it is argued that the esssential elements of this approach could be deployed in other legal systems.

Keywords: duty of care – risk assessment – safety – choice architecture – design – online harms

Chapter 1. Introduction

Early policy-making in the context of the Internet saw the positives of the 'information society' and sought to minimise roadblocks on the 'information superhighway'. The legal framework dealing with 'intermediaries', which remains in place more than two decades later, aimed at removing disincentives to innovation in the sector.[1] A commonality between the EU and American approach was to protect intermediaries from exposure

1 Concerns about innovation remain – see e.g. D. Geradin, "Online Intermediation Platforms and Free Trade Principles: Some Reflections on the Uber Preliminary Ruling Case" in Ortiz (ed), *Internet: Competition and Regulation of Online Platforms*, (Competition Policy International, 2016).

to legal liability in respect of user content hosted or disseminated across their respective services, though the two regimes nonetheless differed in the scope of protection offered. Even by the early 2000's, when fewer people were online and less frequently so, concerns about abuse of the internet were starting to arise. Twenty years on, a wider range of threats are perceived, some arising from specific types of content for example hate speech, others from behaviours, including addiction. Pressure for regulatory action has grown, but much has focussed on dealing with individual items of content and the possibility of removing intermediaries' immunity. This chapter challenges that approach and proposes an alternative approach, based on work done under the aegis of the Carnegie UK Trust, what might be termed a systems-based approach and implemented – in the UK context – by a 'statutory duty of care'.[2] The elaboration of this approach, and the assumptions underpinning it, has the objective of identifying the key elements that could be deployed elsewhere, whether using the same or different implementing mechanisms.

Chapter 2. A Traditional Approach to Liability for Content

Policy in the field of communications, including the mass media, accepted a basic distinction between content creator (including publisher and curator) and those whose role was dissemination – for example, a telecommunications operator. This distinction can be seen, for example, in the development of the EU communications package,[3] though of course there have always been connections between content and network (see e.g. the position of electronic programme guides and the discussion of net neutral-

2 W. Perrin and L. Woods, *'Duty of Care' – Full Report*, April 2019, https://www.carnegieuktrust.org.uk/project/harm-reduction-in-social-media/ developing earlier work in support of a private members bill: https://bills.parliament.uk/bills/1877.

3 This distinction was also present in EU regulation on this issue and can now be found in the European Electronic Communications Code, Directive 2018/1972 , [2018] OJ l 321/36, rec 7; see also views of Court of Justice in Case C-518/11 *UPC Nederland*, judgment 7 November 2013, EU:C:2013:709, para 41; Case C-475/12 *UPC DTH*, judgment 30 April 2014, EU:C:2014:285, para 43; Case C-142/18 *Skype Communications Sarl* v *Institut belge des services postaux et des telecommunications (IBPT)*, judgment 5 June 2019, EU:C:2019:460, para 28. Helberger et al. also note this dichotomy in "Governing online platforms: from contested to cooperative responsibility" (2018) 34(1) *The Information Society* 1-14, p. 2.

ity).[4] A similar concern with the boundary between content creation and curation (ranging from commissioning content, via choices about scheduling and prominence through to ex post moderation) and its dissemination and the role of knowledge in determining the boundary between the two can be seen in the immunity provisions for "information society service" providers in the EU,[5] a distinction implemented in the UK and retained post Brexit. Neutral[6] intermediaries[7] (responsible for transmission, caching or hosting[8]) receive immunity on condition such an intermediary acts expeditiously to remove content once aware of its problematic nature under domestic law.[9] While this frame of analysis may seem appropriate for the transmission infrastructure or for other services that play a purely technical role in the dissemination of bits and bytes, it does not fit so well for some of the online platforms (a term which is only just recently beginning to be defined in legal terms), especially social media platforms which structure to a marked degree the content to which users are exposed. The extent to

---

4 The development of "information society services" (ISS) as a regulatory category blurs this boundary somewhat as they can be content services or more related to transmission; the regulatory response was to carve out some types of ISS from the general regime and treat them as similar to broadcast services: E. Dommering, "General Introduction", in Castendyk, Dommering and Scheuer (eds) *European Media Law* (Alphena/d Rijn: Kluwer Law International, 2008), para 10. See also text attached to n 5 et seq below.

5 Articles 12-14 e-Commerce Directive, Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L178/1

6 Case C-324/09 *L'Oreal v eBay*, [2011] ECR-I 6011 (Grand Chanber), para 124 and see para 122 for examples of activity that a diligent economic operator may engage in; the test of 'diligent economic operator' was applied by the Northern Irish Court of Appeal in *C.G.* v *Facebook Ireland Ltd* [2016] NICA 54, para 72.

7 This has been described as a 'catch-all term': J. Weaver 'Google IP Infringements: No results found?' (2018) 40 EIPR 759; see also M. Husovec, *Injunctions Against Intermediaries in the European Union: Accountable but not liable?* (Cambridge: Cambridge University Press, 2017), 16-17.

8 Originally these phrased were included in Articles 12-14 e-Commerce Directive, but definitions have been expanded in the Proposal for a Regulation on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC (COM/2020/825 final), 15 December 2020; overview of services in scope provided by, for example, A. Vijay, "Liability of internet service providers – a review study from the European perspective" (2019) 41 *EIPR* 451; D. Fernàndez, "ISP Liability Between EU and USA" (2016) 17 *Computer Law Review International* 36.

9 What this means has not yet been fully harmonised: see e.g. Husovec (n 7), pp 52-57.

which those platforms could be said to have knowledge of this content though is open to debate; while the platform processes influence what users see, much of this process is automated.[10]

There has been increasing concern about the availability and prevalence of certain types of content on the Internet, specifically on social media platforms. Concerns about child sexual abuse and exploitation material as well as terrorist content have been a subject of concern since the early 2000's but there are now a wider range of concerns.[11] Solutions have considered making the take-down of content more effective (and solutions in this field would clearly be useful); some have suggested that immunity be removed.[12] Focussing a regulatory regime aimed at platforms on the content they host is, however, problematic. While platforms may prompt or promote certain types of content, they do not create it or commission it; they are not responsible for it in the same way as those that create or reuse that content. Moreover, the size of some of the platforms is in itself an issue; so much content is uploaded (which brings issues of speed as well as of scale) that it would be hard to consider items of content individually (and automated techniques bring their own issues). Moreover, the range of types of content and their audiences are wide and diverse with different expectations in relation to those different types of content. The assessment of the acceptability of items of content is to a large degree context specific. While countries will vary as to their tolerance for certain types of content, speech may be understood differently within those countries or by sub-groups within those countries. Ofcom noted some of these problems given that 'the internet is fundamentally different from television and radio in its nature, audience and scale'.[13] Moreover, this is an area in which there is not only variety in service type but also frequent innovation. Any approach

10 See e.g. Vijay (n 8), p.454; T. Gillespie, *Custodians of the Internet* (New Haven/London: Yale University Press, 2018), p. 7.

11 See issues identified in DCMS, Internet Safety Strategy – Green Paper, October 2017, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650949/Internet_Safety_Strategy_green_paper.pdf.

12 Committee on Standards in Public Life, Intimidation in Public Life: A Review by the Committee on Standards in Public Life (Cm 9543), December 2019, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/666927/6.3637_CO_v6_061217_Web3.1__2_.pdf.

13 Sharon White, "Tackling Online Harm – a regulator's perspective", speech by Sharon White to the Royal Television Society, 18 September 2018, https://www.ofcom.org.uk/about-ofcom/latest/media/speeches/2018/tackling-online-harm (accessed 18 March 2021); see also OFCOM, *Discussion Document: Addressing Harmful Content Online*, p.25.

to regulation would need, therefore, to be to some degree future-proof. As the Interim Report of the DCMS Select Committee fake news inquiry recognised, what is needed is an approach that recognises a 'third way', one that is not dependent on a simplistic division between content on transmission.[14]

Chapter 3. A Different Model

Thinking about social media platforms as quasi-publishers limits the possible policy responses. A different analogy may give rise to different policy options and a return to the language of the 1990's – to "cyberspace"[15] (the virtual world created by the links between computers) – may provide a hint as to where to look for alternative inspiration. The range of services provided across the Internet is wide and may be used differently by different groups; these services provide the place for lots of different activities to happen on-line as take place in a range of spaces off-line. They provide a mechanism for users to engage with one another, to be entertained, to discover information, to advertise and to buy and sell. In the off-line context, providers of spaces are not necessarily regulated in relation to what happens in that place (though some may be – e.g. pubs, casinos, sandwich shops) but they each have some responsibility for the safety of the place, a responsibility which is often dealt with through an assessment of hazards and risks and the likelihoods of harm arising to users of the space. Space management also communicates different expectations as to user behaviour in those spaces. This then leads us to the position that, rather than imposing liability on platforms for individual items of content, they should be expected to assess their respective platforms for safety of their users, and others affected by the service, taking into account how those platforms are used. In moving away from content-focussed regulations, the difficulties in dealing with different understandings about the meaning and acceptability of certain types of content in different jurisdictions, as well as issues arising from scale, may be ameliorated.

14 DCMS Select Committee, *Disinformation and 'fake news': Interim Report* (Fifth Report of Session 2017–19), 24 July 2018 (HC 363).

15 The term is derived from William Gibson's novel *Neuromancer* (Victor Gollancz, 1984).

One might ask, however, what harm may arise from a platform apart from the content itself? The inherent constraints which are found in the physical world do not operate online, and this has allowed the introduction of sophisticated choice architectures aimed at maximising user interaction. This is not necessarily bad, but nor is it neutral – especially when we compare people's interactions online with those offline. It has long been noted that people speaking online experience a disinhibition effect[16] (though the causes are not yet fully understood). Given that users' online experience is mediated by the platforms, the design of the platform could seek to compensate for this; to remind users that others using social media are (in the main) humans too.[17] However, the motivating objective in platform design seems to have been the support of the service providers' bottom line, regardless of consequence. Designing to maximise user engagement for the purpose of acquiring data and delivering adverts, it seems the platforms rather seek to exploit our cognitive weaknesses.[18] So, while a 'like button' can be used as a substitute for nonverbal cues that might be otherwise absent and be seen by the user as a signal of appreciation, for the platforms it is data the accumulation of which can be exploited to understand much more about users than those users may appreciate. A range of adverse consequences has arisen, which some have linked back to design choices, and which risk endangering the well-being of individuals and the functioning of democratic societies: cyber-bullying and hate speech; the polarisation of public debate and the rapid spread of false (and

16 J. Suler, 'The Online disinhibition effect" (2004) 7(3) *Cyberpsychol Behav* 321-6, doi:10.1089/109493104129295.

17 Work on tools and techniques for this is starting in some areas: see e.g. the Prosocial Design Network which lists features and the prosocial consequences they might have and seeks to test them, https://www.prosocialdesign.org/.

18 S. Zuboff, *The Age of Surveillance Capitalism– The Fight for a Human Future at the New Frontier of Power*, (1st ed) (Profile Publishers: London, 2019); in an earlier article she describes "a ubiquitous networked institutional regime that records, modifies, and commodifies everyday experience from toasters to bodies, communication to thought, all with a view to establishing new pathways to monetization and profit" "Big other: surveillance capitalism and the prospects of an information civilization" (2015) 30 *Journal of Information Technology* 75-89, p. 81.

harmful) information.[19] Commentators have pointed to the dangers of content creators responding to the metrics provided by many platforms, whether to sell products themselves (including influencers) or to chase the feel-good glow and being 'liked' – and users are thereby trained to produce response-creating content[20]. Others note that the tools provided for promoting content, aimed at driving user engagement and in effect operating as a trap,[21] priorise extreme, violent and shocking content – that which engages strong negative emotions – with the risk that, for example, conspiracy theories are promoted.[22] Similarly, lies travel faster than the truth (though whether lies are believed is another question);[23] misinformation may thrive because off-line epistemic cues and gatekeeper controls are absent, or because users are nudged to respond and to share or are distracted from considering accuracy.[24] The way information is presented may affect user behaviour: Facebook ran an experiment on its users' newsfeeds that suggested that including social information in an "I voted" button (in this case, displaying faces of friends who had clicked on the button) affected both click rates and real-world voting.[25] Targeted advertising, based on who knows what grounds, raise questions about not

19 S. Bradshaw and P. N. Howard, *The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation* (Working Paper 2019.2: Project on Computational Propaganda) (Oxford, 2019).

20 W. J. Brady et al., "How Social Learnings Amplifies Moral Outrage Expression in Online Social Networks" (2021) (paper under review, available: https://psyarxiv.com/gf7t5/).

21 Anthropological research suggests that those coding recommender algorithms see their function as 'hooking' users; that these algorithms operate as a trap: N. Seaver, "Captivating algorithms: Recommender systems as traps" (2018) Journal of Material Culture, https://journals.sagepub.com/doi/10.1177/1359183518820366.

22 E. Hussein et al., "Measuring misinformation in video search platforms: An audit study on YouTube" (2020) Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1), Article 48. doi 10.1145/3392854.

23 See J. Allen et al., "Evaluating the fake news problem at the scale of the information ecosystem" (2020) 6(14) *Sci Adv* eaay3539, doi: 10.1126/sciadv.aay3539 ; Kozyreva et al., "Citizens versus the Internet: Confronting Digital Challenges with Cognitive Tools" (2020) 21(3) *Psychol Sci Public Interest,* 103-156, doi: 10.1177/1529100620946707.

24 G. Pennycook et al., "Shifting attention to accuracy can reduce misinformation online" (2021) *Nature*, 17 March 2021, https://doi.org/10.1038/s41586-021-03344-2.

25 Kozyreva (n 23).

just manipulation[26] but also intrusion into our respective *fora internum*.[27] Concerns have long been raised about 'filter bubbles' but more generally about the range of topics of information users receive.[28] It has also been suggested that the very short-form format of news based on headlines and snippets gives users the illusion of being informed.[29] Targeting may be weaponised by nefarious actors.[30] 'Sock puppet accounts' and networks of coordinated accounts may spread and embed false information and sow discord. While users are not just passive recipients in the online environment,[31] and users may innovate and disrupt at least some of the time, it must be recognised that not everybody has the capability to hack the system. As Leiser notes, some of the theoretical models in this area have fallen into a common trap: that of assuming that all users are rational and fully informed; and underplaying the role of cognitive weaknesses most humans exhibit.[32] Additionally, the tools provided to users to take control

---

26 S. Matz et al., "Psychological targeting as an effective approach to digital mass persuasion" (2017) 114(48) *Proc Natl Acad Sci USA* 12714, doi: 10.1073/pnas.1710966114.

27 S. Alegre, "Rethinking the Right to Freedom of Thought in the 21st Century" (2017) 3 *Eur. Hum. Rights. Rev* 221; S. Zuboff (n. 11); S. Alegre, "Regulating around Freedom I the "forum internum"" (2021) *ERA Forum* 591.

28 C. Sunstein, "Republic.com 2.0", p. 5; in *#Republic: Divided Democracy in the Age of Social Medi*a (Princeton, NJ, USA, and Oxford, UK: Princeton University Press, 2017), Sunstein also notes 'asymmetrical updating', that is a strong tendency to favour evidence that confirms our beliefs and ignore or misread evidence that does not. How to compensate for this does not seem to be a simple matter of ensuring more diverse viewpoints are presented. While some studies (e.g. Bakshy et al., "Exposure to ideologically diverse news and opinion on Facebook" (2015) 348 *Science* 1130, DOI 1-.1126/science.aaa1160) suggest that user choice may be part of this, others have suggested that algorithmic amplification has a role to play through the creation of a variant of feedback loop: A. J. B. Chaney et al., "How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility" (2018) RecSys '18, October 2–7, https://arxiv.org/pdf/1710.11214.pdf.

29 S. Schäfer, "Illusion of knowledge through Facebook news? Effects of snack news in a news feed on perceived knowledge, attitude strength, and willingness for discussions" (2020) 103 *Computers in Human Behavior* 1–12. 10.1016/j.chb.2019.08.031.

30 See concerns expressed by the DCMS Select Committee, *Disinformation and 'fake news': Final Report*, 18 February 2019, https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/179102.htm.

31 A. Murray, *Regulation of Cyberspace*, (2007, Oxford University Press).

32 M. Leiser, "The Problem with 'Dots': questioning the role of rationality in the online environment" (2016) 30 *International Review of Law, Computers and Technology* 191.

of their own environment are not extensive and may not be easy to use nor recognise specific risks or problems faced by particular groups.

It is the fact that these choices lie in the hands of the operators meaning that placing responsibility on the operators for the design and operation of their respective platforms is legitimate; they are being held responsible for their own actions, not those of others. The designers are the risk-creators and thus best-place to manage those risks.[33] While not all the possible issues are fully understood, platform operators can still ask themselves the question how this service is working; is there evidence that there might be side effects; what content and safety curation tools can we provide (especially considering some groups may have particular needs); and what the alternative to a given feature is? Perhaps all inventors and designers should ask themselves, 'what happens when this scales and what happens when the bad people get hold of it?' In this, the approach looks at features and user behaviours and their likely impacts at a general level, not assessing individual items of content.

Chapter 5. Risk Assessment: A Model from Work Spaces

If we think of social media platforms as quasi-public spaces, the regulation ensuring those spaces are safe may constitute a model for the implementation of the system-based approach. In the UK, the main mechanism is found in the Health and Safety at Work Act 1974 (HSWA).[34] It provides a statutory duty of care – that is a duty of care similar to that found in the tortious doctrine of negligence – but specified (and possibly amended) by the terms of legislation. Section 2(1) HSWA states:

> It shall be the duty of every employer to ensure, so far as is reasonably practicable, the health, safety and welfare at work of all his employees.

This is a very broad category and s. 3(1) extends the duty beyond the employer's duty to employees to include "persons not in his employment who may be affected" by the business. The Act also imposes reciprocal duties on the employees.

---

33 Robens Report: Safety and Health at Work, July 1972 (Cmnd 5034).

34 For the development of the statutory duty of care and its difference from the duty of care found in the common law doctrine of negligence see L. Woods, "The duty of care in the Online Harms White Paper" (2019) 11(1) *Journal of Media Law* 6.

While the nature of the obligation is broad – in the case of the duty to employees, it is to prevent harm and as regards others it is avoidance of exposure to risks to their health or safety – the HSWA gives examples of specific issues about which the employee must take action. Examples include: provision of machinery that is safe; the training of relevant individuals; and the maintenance of a safe working environment. This list of actions does not replace the general duty. The HSWA additionally contains an obligation on an employer "to prepare and as often as may be appropriate revise a written statement of his general policy with respect to the health and safety at work": this is the beginnings of formalising a preventative approach, based on an assessment of risks posed.

The regime is enforced by a regulator, the Health and Safety Executive, which has a range of powers including "improvement notices", "prohibition notices" and prosecution. Recourse to the criminal law is a matter of last resort and sentencing guidelines identify factors that influence the heaviness of the penalty. Factors that tend towards high penalties include flagrant disregard of the law, failing to adopt measures that are recognised standards, failing to respond to concerns, or to change/review systems following a prior incident as well as serious or systematic failure within the organisation to address risk. So, while the duty of care is still described as being owed to a certain group of people (employees in s. 2(1) and persons "affected by an undertaking" in s. 3(1)), general enforcement powers lie elsewhere. Individuals suffering injury are not empowered to bring action under this regime; injury suffered is dealt with through traditional negligence claims. This point highlights the difference between individual instances of harm and the environment giving rise to the risk of harm.

There are a number of points which suggest that an over-arching duty such as that found in HSWA is an appropriate model. It applies widely and in a range of different sorts of contexts; it applies to almost all employers and the myriad activities that go on in them. A similar tool could presumably be deployed across social media and the many purposes for and ways in which those platforms are used. A factor in the general duty's usefulness is the fact that, with the exception of a limited number of high risk activities which are controlled by specific regulations[35], it does not set down detailed rules with regards to what must be done in each workplace. It rather sets out some general duties that employers have both as regards their employees and the general public, but leaves the employer

35 For example, see Control of Major Accident Hazards Regulations 1999 (SI 1999/743).

to identify appropriate implementation mechanisms. This allows the employer's obligation to be tailored to the specific risks found in a particular (work) environment, subject to the guidance from the regulator. As well as providing for flexibility within the current range of providers, it allows a certain degree of future-proofing as new features, services or problems are introduced. It also allows for new research on understanding risks and how to mitigate against them to be taken into account as that body of research develops. An outcome orientated approach, which implies that an employer should seek to identify steps that would be reasonably effective in the relevant context, also mitigates the risk of a tick box approach were specific, detailed rules (e.g. ban bots; prohibit anonymous accounts) to be adopted. Finally, the distinction between the environment creating the risk of harm and the individual instances of harm broadly parallels the distinction between the systems constituting the platform/service and the individual instances of content or behaviour.

Chapter 6. The Statutory Duty of Care: A Proposal

This leads us to system-based regulation, where 'system' is understood in two ways:

- the focus of regulation is on the software system (or more broadly the service, including the business model) itself rather than on the content hosted on the service; and
- providers of such services should have a system (understood as a process) in place to risk assess the service and individual features of the service – and to take appropriate steps to address concerns arising.

The operator of the system should be subject to an overarching, general duty of care. The duty of care must set out the persons to whom the duty is owed,[36] the types of harm from which that person should be protected as well as the operators within scope.

As regards the first point, the Carnegie proposal suggested that both users and non-users of a service were owed a duty of care, provided that non-users were affected by the operation of the platform. In this, it followed the model of the HSWA. The reasoning was that persons could be

36 Note the Environmental Protection Act 1990 uses a similar mechanism but does not identify the beneficiary of the duty.

harmed by behaviours on a platform even if they had not joined it, for example in the case of "revenge porn".

The proposal also noted that it was important that the types of harm be identified in statute,[37] but that the vectors of harm may be elaborated in regulatory guidance (especially in the light of developing research). Although the types of harm need some clarification, these can be reasonably broad categories, as the HSWA demonstrates; regulatory guidance can fill in the details. These categories of harm should be identified by reference to the impact on the victim, not by reference to whether the speech might be considered illegal or not.[38] The criminal law is not always the best proxy for understanding harm and, crucially, also does not focus on the role of the platform itself in encouraging, facilitating or exacerbating the occurrence of harm. As noted above, it is the fact that the platforms are risk creators that justifies the decision to regulate at this point.

The Carnegie proposal sought to define social media, on the basis of the following characteristics – that services:

- have a strong two-way or multiway communications component;
- display user-generated content;[39]
- publicly or to a large member/user audience or group.

This could include some private messaging apps that allowed large groups to communicate. Search engines were excluded because, although they have an effect on the information provided to users, they may give rise to issues surrounding the right to information, prominence and diversity which may necessitate a different response. Also excluded were actors,

37 See similarly Digital, Culture, Media and Sport Select Committee, *Disinformation and 'fake news': Final Report*, Eighth Report of Session 2017-19 (HC 1791), 18th February 2019, paras 31-32; in other sectors, e.g. broadcasting as well as the HSWA, regulators are entrusted with understanding the precise meaning of harm.

38 In this, the proposal differs from the characteristics that Cole, Etteldorf and Ullrich ascribe to duty of care models: Cole, Etteldorf and Ullrich, *Cross-border Dissemination of Online Content* (Baden-Baden: Nomos Verlagsgesellschaft, 2020), p 202 which limits risk assessment to illegal content and behaviours.

39 The Audiovisual Media Services refers to use-generated content and contains a definition of "user generated video"; the UK implementation of this provision does not use the same terminology. For discussion of the difficulties with the definitions in the Audiovisual Media Services Directive see L. Woods, "Video-sharing platforms in the revised Audiovisual Media Services Directive" (2018) 23 (3) *Communications Law* 127.

essentially the broadcast and print media that are already subject to regulatory or self-regulatory regimes.

The essential element of this model is a risk assessment considering the service, including its individual features, and the business model of the service. The focus of enquiry is the impact the structures and business choices have in creating a risky environment. The system-based approach is neutral as to the topics of content (though part of that system will involve dealing with complaints and with content that is contrary to the law); as such, the system may be less open to the accusation that regulation will result in excessive take-down on unclear bases[40].

Risk assessments require the identification of hazards (that is something that could cause harm) and determine how likely it is that each hazard will occur and how severe the consequences would be. A risk assessment should take into account relevant human rights. Freedom of expression is obviously important but it is not the only right. Moreover, design choices may have discriminatory effects in the enjoyment of rights (the use of AI in content moderation is one example). The assessment of consequences operates at a general level rather than seeking to determine outcomes in particular cases. In this there is a difference from a regime aimed at compensating individual victims. The starting point is the platform and the likely consequences of its use; it is not about starting with an instance of harm or a category of content and trying to work backwards in respect of that particular example. As a final stage, the operator should determine the appropriate mitigating steps – whether this be not to deploy the new feature/change, to amend it, or to bring in some compensating measure. At the least, the operator should perform risk assessment before introducing new processes or activities, before introducing changes to existing processes or activities (such as a significant change to an algorithm), or when the company identifies a new hazard (e.g. becomes aware of research); it should also monitor whether the mitigating steps seem to be effective. This process was described as instituting a harm reduction cycle. We envisaged that a regulator would have some say in identifying what a good risk assessment looks like, but for risky services (including large services), the Carnegie proposal also envisaged some involvement of relevant civil society actors. In this, transparency at some level of granularity and within a framework set by the regulator, is key.

The duty is not focussed on particular technologies or the problems they cause. It allows a platform to take into account the interplay of

40 Cole et al (n 38) note this criticism, p. 204.

different features in terms of risk assessment and mitigation. It is also not limited to technical specifications, but may take into account when, how and to whom features or services are deployed.[41] As HSWA illustrates, the fact that the statutory duty of care is a general obligation does not mean that statute cannot specify specific obligations within that general duty – for example, the need to have an effective complaints mechanism, obligations of transparency for particular issues, the need to take particular steps with regard to specific types of content (e.g. child sexual abuse and exploitation material).

In carrying out their duty of care, platform operators are not expected to achieve perfection. An appropriate threshold is similar to that found in the doctrine of negligence; it is not a strict liability regime. Rather, an operator should take reasonable steps in relation to foreseeable harms. Whether an operator has satisfied the duty will be determined by the regulator; jurisprudence from the doctrine of negligence is not binding in this regard.[42] "Reasonable" and "foreseeable" should take into account the platform's use, including its user base size and profile, as well as any relevant industry standards. While the service provider may not engage in wilful blindness, nor should they be judged with the benefit of hindsight. "Reasonable steps" do not require a perfectly sanitised environment; rather the requirement aims to consider the role platforms play in creating or exacerbating the problems. Moreover, the mere fact that there problematic content or behaviours may be found on a platform does not in and of itself constitute a violation of the duty of care. Ultimately, while the regime is orientated towards a particular result, the question of whether an operator has satisfied its duty of care is not answered by numbers of take-downs nor numbers of problematic posts/instances of use (though a platform on which there are many instances of problematic content may be less likely to have satisfied the duty of care). Liability is about engagement which the risk assessment and mitigation process; it does not involve liability for content.

As a result of the focus on design, the tools and changes are not limited to ensuring that a take-down regime operates effectively and fairly, though it should do that. There are three main points of influence before we reach the question of whether content should be taken down: the point at which a user engages with the platform (including sign up processes, means of

41 In this it is different from proposals which focus on a specific technology or technical standard, outlined Cole et al (n 38), p. 202.

42 Clerk and Lindsall on Torts (23rd ed) para 8-56.

finding others in a group, and tools to communicate for example augmented reality filters/overlays); the mechanisms by which content is disseminated (e.g. search engines, hashtags, recommender systems, newsfeeds); and the mechanisms by which recipient users engage with content, including choosing not to engage with it, but also mechanisms such as tools for sharing/forwarding/demonstrating approval or disapproval. Examples of this category include retweeting, liking, forwarding tools, as well as those allowing users to block or mute incoming messages. Each of these points may have an impact on the content available – in terms the content created as well as the way content flows across platforms. Significantly, as many interventions allow speech to continue, they may be less intrusive to users' freedom of expression.[43]

Insofar as platforms operate as advertising services, the duty of care should extend to this aspect of the service too, with regard to protecting users.[44] Questions that might be asked include whether the platform engages in any KYC ("know your client") processes as regards advertisers; and what sorts of ads does it permit – do any require specific safeguards? Further, how are audiences segmented (e.g. what controls are there around permitted groupings/topics – are any segments impermissible or undesirable)? The availability of micro-targeting itself should be assessed for its risks.

The last port of call is take-down. An operator needs to ensure that it has an adequate complaints mechanism that is accessible and easy to use and which operates in a fair, timely and transparent manner.[45] As well as reporting on numbers and speed of take-down, reporting should consider what is being taken down, and why, as well as categories of complainant (with the intention of not only identifying where unforeseen problems arise, but also identifying and mitigating against discrimination in the complaints system).

---

43 For a consideration of the issues and some of the difficulties surrounding this analysis in the context of the Carnegie UK Trust proposal, see L. Woods, "The Carnegie Statutory Duty of Care and Fundamental Freedoms", 2019, https://www.carnegieuktrust.org.uk/publications/doc-fundamental-freedoms/.

44 This viewpoint was adopted by the Centre for Data Ethics and Innovation in its recommendations to Government: CDEI, *Review of Online Targeting*, 4 February 2020, https://www.gov.uk/government/publications/cdei-review-of-online-targeting/online-targeting-final-report-and-recommendations.

45 The House of Lords Communications Committee noted the need for consistent enforcement as well as transparency of complaints handling: *Growing up with the Internet* (2nd Report of Session 2016–17) (HL Paper 130), 21 March 2017, paras 241-2.

The Carnegie proposal also envisaged that platforms should develop a triage process for emergent problems; while the detail of the problem may be unknown, it is fairly certain that new problems will arise, as the issue of misinformation and disinformation related to Covid-19 illustrates. The interface with law enforcement and relevant regulatory authorities (e.g. Advertising Standards Authority, Financial Conduct Authority) in the exercise of their powers should also be considered.[46]

The increasingly problematic nature of the social media environment suggests that self-regulation (even self-regulation engaging with voluntary codes of practice) has not worked well. Moreover, reliance on users to take action before the courts is unlikely to constitute a sufficient corrective for a range of reasons but notably because of the asymmetry of resources and knowledge between the major platforms and litigants. A regulator is required, even if the proposed scheme is not a traditional top-down command scheme. It is crucial, especially given the importance of freedom of expression in the functioning of a democracy, that the regulatory be independent from both industry and from government. It must make decisions based on objective evidence (and not under pressure from other interests) and be viewed as a credible regulator by the public. Independence means that it must have sufficient resources, as well as relevant expertise. A completely new regulator created by statute would take some years before it was operational. The Carnegie proposal therefore envisaged extending the powers of the existing telecommunications and media regulator, Ofcom. This approach has a number of advantages. 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.

The responsibilities of the regulator would include identifying actors in scope; developing good practice and guidance about harms and vectors by which harm could be caused (including where appropriate approving industry codes of practice and standards); monitoring the harm reduction cycle and risk assessment processes; and enforcing the duty of care. The

46 The obligations on platforms to cooperate have arising in the enforcement of intellectual property rights, especially in connection with loss of immunity; see e.g. Husovec (n 7). Cooperation with regulatory authorities and law enforcement has drawn less attention, but see e.g. mechanisms envisaged by the recently agreed Regulation on addressing the dissemination of terrorist content online: Regulation 2021/784 [2021] OJ L172/79. As part of the Carnegie Proposal a model was proposed: see W. Perrin and L. Woods, "Online Harms – Interlocking Regulation" (Blog), 11 September 2020, https://www.carnegieuktrust.org.uk/blog/online-harms-interlocking-regulation/.

Carnegie proposal also included information gathering powers for the regulator.[47] As in many other regulatory fields, failure to comply should be a violation of the regime in and of itself.

Finally, the regime must have sanctions, though any enforcement action should be context specific and proportionate, especially given the fundamental rights in play (including but not limited to freedom of expression). The range of mechanisms available within the HSWA are interesting because they allow the regulator to try improve conditions rather than just punish the operator; to some extent the GDPR and the Data Protection Act 2018 have a similar approach. Other options include adverse publicity orders where the operator is required to display a message on its screen most visible to all users detailing its offence which could result in reputational losses.[48] Another possibility, albeit one that would require some thought in terms of implementation, is borrowing techniques from restorative justice.[49] For those that will not comply, the regulator should be empowered to impose fines, including GDPR or competition policy magnitude fines. The more difficult questions relate to what to do in extreme cases. Should there be a power to send a social media services company director to prison (as in the HSWA) or to turn off the service? The Digital Economy Act 2017 (DEA) contains power[50] (which was never brought into force) for the age verification regulator to issue a notice to internet service providers to block a website in the UK. Blocking orders, even if technically effective, raise concerns about 'collateral censorship' – where a platform is blocked the speech rights of the platform's users are affected. This is particularly the case where there are large platforms carrying many different types of content (most of which would be unproblematic). These sorts of mechanisms – as well as criminal sanctions for speech – raise questions about their proportionality from a freedom of expression perspective. The DEA provided what could be a middle ground, though again this provision has not been brought into force. Section 21 empowers

47 On the importance of evidence gathering powers, see the evidence of Sharon White to the DCMS Select Committee, *Disinformation and 'fake news': Final Report,* (Eighth Report of Session 2017-19) (HC 1791), 18 February 2019, para 33.

48 On the effectiveness of mechanisms leading to reputational loss see e.g. Armour et al., "Regulatory Sanctions and Reputational Damage in Financial Markets" (2017) 52(4) *Journal and Financial and Quantitative Analysis* 1429 – 1448.

49 Restorative justice is used in the context of criminal justice in England and Wales; see here for CPS guidance: https://www.cps.gov.uk/legal-guidance/restorative-justice.

50 Section 23 Digital Economy Act 2017.

the regulator to issue notices to others who are dealing with the non-complying operator, such as credit card or other payment services. According to the Explanatory Memorandum to the DEA, the purpose of such a notice is to bring the problem to the attention of these ancillary service providers so as "to enable them to consider whether to withdraw services",[51] thus disrupting the provision of the service. This approach might be deemed problematic in that it uses private actors as enforcement mechanisms,[52] though it should be noted that similar techniques have been used in other regulatory contexts (e.g. cinemas were used as enforcement mechanisms for age ratings for films).

Chapter 7. Conclusion

This paper has sought to distinguish between two models of regulation in respect of social media: that aimed at content, which has been traditionally used in the context of speech concerns and specifically in relation to the mass media; and systemic regulation, which takes a process-based risk assessment approach to regulation used in many industrial sectors. Drawing on insights about the impact of design and choice architecture on user freedom and behaviour, and based on the work of Carnegie UK Trust, it has argued for the target of regulation to be the software and business systems that make up social media services. Not only do these systems have an impact on user behaviour but choices about the design and deployment of such systems are under control of the relevant companies. Looking to the UK legal environment, Carnegie UK Trust proposed a particular vehicle by which systemic regulation could be deployed: the statutory duty of care to create a general obligation enforced by a regulator rather than ex post individual litigation. While the statutory duty of care as a vehicle

51 Explanatory Memorandum to the Draft Online Pornography (Commercial Basis) Regulations 2018, available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/749750/Explanatory_Memorandum_to_the_Draft_Online_Pornography__Commercial_Basis__Regulations_2018.pdf

52 The DEA did not impose penalties on those which did not cooperate; in this it might be different from the context of intermediaries in intellectual property. More generally see M. MacCarthy, "What Payment Intermediaries are Doing about Online Liability and Why it Matters" (2010) 25 *Berkley Technology Law Journal* 1037, especially p 1056.

to implement this model may be particular to the UK, the underlying regulatory model could be deployed in other jurisdictions.

Bibliography

Alegre, S. "Rethinking the Right to Freedom of Thought in the 21st Century". (2017) 3 *Eur. Hum. Rights. Rev* 221.

Alegre, S. "Regulating around Freedom I the 'forum internum'". (2021) *ERA Forum* 591.

Allen, J. et al. "Evaluating the fake news problem at the scale of the information ecosystem". (2020) 6(14) *Sci Adv* eaay3539, doi: 10.1126/sciadv.aay3539A.

Armour et al. "Regulatory Sanctions and Reputational Damage in Financial Markets." (2017) 52(4) *Journal and Financial and Quantitative Analysis* 1429 – 1448.

Bakshy et al. "Exposure to ideologically diverse news and opinion on Facebook". (2015) 348 *Science* 1130, DOI 1-.1126/science.aaa1160.

Bradshaw, S and Howard, P. N. *The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation.* (Working Paper 2019.2: Project on Computational Propaganda) (Oxford, 2019).

Brady, W. J. et al. "How Social Learnings Amplifies Moral Outrage Expression in Online Social Networks" (2021). (paper under review, https://psyarxiv.com/gf7t5/).

CDEI. *Review of Online Targeting*. 4 February 2020. https://www.gov.uk/government/publications/cdei-review-of-online-targeting/online-targeting-final-report-and-recommendations.

Chaney, A. J. B. et al. "How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility." (2018) RecSys '18, October 2–7. https://arxiv.org/pdf/1710.11214.pdf.

Cole, Etteldorf and Ullrich. *Cross-border Dissemination of Online Content.* Baden-Baden: Nomos Verlagsgesellschaft, 2020.

Committee on Standards in Public Life. Intimidation in Public Life: A Review by the Committee on Standards in Public Life (Cm 9543). December 2019. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/666927/6.3637_CO_v6_061217_Web3.1__2_.pdf.

DCMS. Internet Safety Strategy – Green Paper. October 2017. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650949/Internet_Safety_Strategy_green_paper.pdf.

DCMS Select Committee. *Disinformation and 'fake news': Interim Report* (Fifth Report of Session 2017-19). 24 July 2018 (HC 363).

DCMS Select Committee. *Disinformation and 'fake news': Final Report***.** Eighth Report of Session 2017- 19 (HC 1791). 18 February 2019. https://www.parliament.uk/business/committees/committees-a-z/commons-select/digital-culture-media-and-sport-committee/news/fake-news-report-published-17-19/.

Dommering, E. "General Introduction". In: Castendyk, Dommering and Scheuer (eds). *European Media Law.* (Alphena/d Rijn: Kluwer Law International, 2008).

Fernàndez, D. "ISP Liability Between EU and USA". (2016) 17 *Computer Law Review International* 36.

Geradin, D. "Online Intermediation Plattforms and Free Trade Principles: Some Reflections on the Uber Preliminary Ruling Case." In: Ortiz (ed). *Internet: Competition and Regulation of Online Platforms*. (Competition Policy International, 2016).

Gillespie, T. *Custodians of the Internet.* (New Haven/London: Yale University Press, 2018).

Helberger et al. "Governing online platforms: from contested to cooperative responsibility". 2018.

House of Lords Communications Committee. *Growing up with the Internet* (2nd Report of Session 2016–17) (HL Paper 130). 21 March 2017.

Husovec, M. *Injunctions Against Intermediaries in the European Union: Accountable but not liable?.* (Cambridge: Cambridge University Press, 2017).

Hussein, E. et al. "Measuring misinformation in video search platforms: An audit study on YouTube". (2020) Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1), Article 48. doi 10.1145/3392854.

Kozyreva et al. "Citizens versus the Internet: Confronting Digital Challenges with Cognitive Tools". (2020) 21(3) *Psychol Sci Public Interest* 103-156, doi: 10.1177/1529100620946707.

Leiser, M. "The Problem with 'Dots': questioning the role of rationality in the online environment." (2016) 30 *International Review of Law, Computers and Technology* 191.

MacCarthy, M. "What Payment Intermediaries are Doing about Online Liability and Why it Matters." (2010) 25 B*erkley Technology Law Journal* 1037.

Matz, S. et al. "Psychological targeting as an effective approach to digital mass persuasion". (2017) 114(48) *Proc Natl Acad Sci USA* 12714, doi: 10.1073/pnas.1710966114.

Murray, A. *Regulation of Cyberspace*. (2007, Oxford University Press).

OFCOM. *Discussion Document:Addressing Harmful Content Online.*

Pennycook, G. et al. "Shifting attention to accuracy can reduce misinformation online". (2021) *Nature*, 17 March 2021. https://doi.org/10.1038/s41586-021-03344-2.

Perrin, W and Woods, Lorna. *"Duty of Care" – Full Report*. April 2019. https://www.carnegieuktrust.org.uk/project/harm-reduction-in-social-media/.

Perrin, W and Woods, Lorna. Online Harms – Interlocking Regulation (Blog). 11 September 2020. https://www.carnegieuktrust.org.uk/blog/online-harms-interlocking-regulation/.

Robens Report: Safety and Health at Work, July 1972 (Cmnd 5034).

Schäfer, S. "Illusion of knowledge through Facebook news? Effects of snack news in a news feed on perceived knowledge, attitude strength, and willingness for discussions." (2020) 103 *Computers in Human Behavior* 1–12. 10.1016/j.chb.2019.08.031.

Seaver, N. "Captivating algorithms: Recommender systems as traps". (2018) Journal of Material Culture. https://journals.sagepub.com/doi/10.1177/1359183518820366.

Suler, J. "The Online disinhibition effect". 2004 7(3) *Cyberpsychol Behav* 321-6, doi: 10.1089/109493104129295.

Sunstein, C. *Republic.com 2.0* (Princeton, NJ Princeton University Press, 2007).

Sunstein, C. *#Republic: Divided Democracy in the Age of Social Media* (Princeton, NJ, USA, and Oxford, UK: Princeton University Press, 2017).

Vijay, A. "Liability of internet service providers – a review study from the European perspective". (2019).

Woods, Lorna. "Video-sharing platforms in the revised Audiovisual Media Services Directive." (2018) 23 (3) *Communications Law* 127.

Woods, Lorna. The Carnegie Statutory Duty of Care and Fundamental Freedoms. 2019. https://www.carnegieuktrust.org.uk/publications/doc-fundamental-freedoms/.

Woods, Lorna. "The duty of care in the Online Harms White Paper." (2019) 11(1) *Journal of Media Law* 6.

Zuboff, S. "Big other: surveillance capitalism and the prospects of an information civilization." (2015) 30 *Journal of Information Technology* 75-89.

Zuboff, S. *The Age of Surveillance Capitalism– The Fight for a Human Future at the New Frontier of Power* (1st ed). (Profile Publishers: London, 2019).