King's Student Law Review

Title: **To what extent is the new regulatory regime proposed by the Online Harms White Paper effective in increasing accountability on Social Media Companies for online hate speech?**

Author: *Osama Shaaban*

Source: *The King's Student Law Review*, Vol XI, Issue II

Cite as: Osama Shaaban (2021) ***To what extent is the new regulatory regime proposed by the Online Harms White Paper effective in increasing accountability on Social Media Companies for online hate speech?*** King's Student Law Review Vol XI Issue II pp 120-138

Published by: King's College London on behalf of The King's Student Law Review

---

To what extent is the new regulatory regime proposed by the Online Harms White Paper effective in increasing accountability on Social Media Companies for online hate speech?

Osama Shaaban[1]

ABSTRACT

The appointment of Ofcom as the regulator for online harms raises a plethora of regulatory concerns for social media companies. Nevertheless, it remains undoubtedly clear that a core concern emanating from the new regulatory regime relates to the extent to which social media companies will be held to higher standards of accountability. Through evaluating the legal mechanisms entitling online hate speech victims to judicial redress prior to appointing Ofcom as a regulator, the paper develops a more holistic understanding of the socio-legal implications that arise following Ofcom's appointment and whether this will sufficiently shield victims from hateful content. Upon establishing the contours of the new regulatory regime, the paper evaluates the excessively wide and nebulous duty of care proposed through the White Paper and recommends a more narrowly delineated duty in context with contractual mechanisms.

I. Introduction

The propensity for social media platforms to disseminate online hate speech has increased exponentially over the past decade. According to the Oxford Internet Institute, as of 2019, a staggering 40% of the United Kingdom **("UK")** population has been exposed to online hate, ranging from online abuse to cyberbullying.[2] However, the existing legal framework

[1] *LPC LLM student at the University of Law*

[2] Olga Robinson, Alistair Coleman and Shayan Sardarizadeh, 'A Report of Anti-Disinformation Initiatives' (*Oxford Internet Institute and University of Oxford*, August 2019) <https://oxtec.oii.ox.ac.uk/wp-content/uploads/sites/115/2019/08/OxTEC-Anti-Disinformation-Initiatives-1.pdf> accessed 27 February 2021

evidences an inadequate enforcement against hate speech; predominantly because defamation law in the UK only captures one form of what might be categorised as hate speech, whilst the judicial interpretation of the E-commerce Directive[3] establishes an elusive standard of what could retrospectively constitute actual knowledge of harmful content.

There are three distinctive features of hate speech on social media companies that make them significantly difficult to regulate: publicness, permanence and reach.[4] Live video streams of the Christchurch mass shootings in New Zealand have shown that the wide reach of online harms can make take-down redundant given the substantial harm individuals face emotionally and mentally. This exemplifies how the publicness of social media companies tends to magnify the effect of hate speech on victims due to the permanence of posts which can initiate severe harm on an individual and collective level if not removed expeditiously.

The Online Harms White Paper evidences the Government of United Kingdoms' **("HM Government")** first attempt to regulate a range of pervasive harms within the online sphere, including hate speech. Whilst receiving full governmental response as of the 15th of December 2020,[5] the White Paper recommends that online harms (including hate speech) are addressed through a co-regulatory model, where the Office of Communications **("Ofcom"**), a telecommunications regulator in the UK could be assigned with the responsibility of imposing obligations on social media companies and issuing penalties if compliance was not exercised.[6] The proposed reform in the White Paper has culminated in parliament drafting the Online Safety Bill which has undergone its first reading in the House of Lords.[7]

[3] Council Directive (EC) 2000/31 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (Directive on electronic commerce) [2000] OJ L 178/1 (E-Commerce Directive).
[4] Chara Bakalis 'Rethinking cyberhate laws', (2018) 27 (1) Information and Communications Technology Law, 104 <https://doi.org/10.1080/13600834.2017.1393934> accessed 12 November 2020.
[5] Department for Digital, Culture, Media and Sport and Home Department, *Online Harms White Paper: Full government response to the consultation* (White Paper, Cm 354, 2020) ch 3. (Online Harms White Paper).
[6] ibid, para 3.5 and 3.6.
[7] Draft Online Safety Bill, May 2021, CP 405.

This paper, however, focuses on critiquing the Online Harms White Paper rather than the proposed Online Safety Bill. This is due to the White Paper clarifying the underpinning rationale behind the proposed Bill, thus adding depth in understanding the context and underlying factors behind the new regulatory regime. The paper specifically focuses on analysing a fundamental pillar outlined in the White Paper; accountability. It evaluates the extent to which the new regulatory regime proposed by the Online Harms White Paper is effective in increasing the threshold of accountability of social media companies (SMCs) for online hate speech.

It is essential to note that this essay is not concerned with criminalising hate speech, but rather concerned with assessing how SMC's behaviour can be addressed from a regulatory perspective. The rationale behind this is that the legal discourse has been majorly oriented towards criminalising hate speech by way of prosecuting individuals rather than focusing on alternative methods of accountability that are more responsive to the victim's true needs. These needs are centred on preventing reputational and emotional harm by removing posts rather than punishing the perpetrator for his harmful actions. [8]

II. Defining Online Hate Speech

To effectively analyse the accountability framework in the Online Harms White Paper, it is imperative to delineate the contours of hate speech in UK legislation. As emphasized by the Alan Turing Institute, there is a lack of consensus regarding the definition of hate speech.[9] A range of legislation have been enacted in the UK for hate crime, however this is distinct from hate speech. This includes the Criminal Justice Act 2003,[10] which criminalises hate speech on the basis of derogatory statements targeting an individual's protected characteristic such as

[8] Frederick M. Lawrence, 'The Punishment of Hate: Toward a Normative Theory of Bias-Motivated Crimes' (1994) 93(2) Michigan Law Review <https://www.jstor.org/stable/1289930?seq=1> accessed 27 January 2021.
[9] Bertie Vigden, Emily Burden and Helen Margetts, 'Understanding Online Hate, VSP Regulation and the Broader Context' (*Alan Turing Institute*, February 2021), p. 2 <https://www.ofcom.org.uk/__data/assets/pdf_file/0022/216490/alan-turing-institute-report-understanding-online-hate.pdf> accessed 3 August 2021.
[10] Criminal Justice Act 2003, s 145, 146.

religion, race or sexual orientation.[11] What ought to be recognised when it comes to hate crime is that they deal with hostile behaviour that is already criminalised under the law (e.g. assault, harassment), that can be aggravated by expressing hate towards a victim's protected characteristic. [12] Hate speech, however – as emphasized by Bakalis, demarcates the illegality of content not only within the premise of protected characteristics, but also to more nuanced and broader forms of hostility and hatred.[13] Here, two points must be highlighted. First, is the fact that cross-fertilising understandings of hate crime to hate speech is not an effective way to create an efficient regulatory framework, as illegality must not correlate with criminality, but with broader forms of hate. The second premise is that offences in hate crime legislations apply both to an online and offline context, which does not take into account distinctive features in online hate (such as publicness, permanence and reach).

Another issue arises is the fact that neither the Online Harms White paper nor the proposed Online Safety Bill have categorically defined 'online hate speech'. Hence, to pave an analytical framework for this paper, Alan Turing's definition of online hate speech will be adopted:

"A communication on the internet which expresses prejudice against an identity. It can take the form of derogatory, demonising and dehumanising statements, threats, identity-based insults, pejorative terms and slurs". [14]

The term 'identity' in the definition above, nevertheless, is shrouded with ambiguity. According to Fearon, identity does not only connote social or religious constructs (i.e., attachment to a country/religion), but also personal beliefs.[15] This highlights the complexity

[11] Law Commission, *Reforms to Hate crime laws to make them fairer, and to protect women for the first time* (Law Com No 250, 2020).
[12] Protection From Harassment Act 1997, s 2.
[13] Chara Bakalis and Julia Hornle, 'The role of social media companies in the regulation of online hate speech' [2019] 85 Studies in Law, Politics, and Society p. 5 <https://qmro.qmul.ac.uk/xmlui/handle/123456789/70038> accessed 17 January 2021.
[14] Vigden, Burden and Margetts (n 8), p. 2.
[15] James D. Fearon, 'What is identity (as we now use the word)?' (*Stanford University*, 1 November 1999) <https://web.stanford.edu/group/fearon-research/cgi-bin/wordpress/wp-content/uploads/2013/10/What-is-Identity-as-we-now-use-the-word-.pdf> accessed 4 August 2021.

of defining hate speech, but the generic definition suffices to analyse the regulatory framework of the Online Harms White Paper with considerably more nuance and depth.

III. Current Legal Framework

To holistically evaluate the efficacy of the proposed regulatory regime by the White Paper, judicial redress mechanisms available to victims of hate speech prior to appointing Ofcom as a regulator must be deliberated upon. These include bringing a civil claim against:

(a) The perpetrator for disseminating online hate
(b) Social media companies for failing to take down posts that contain hateful content.

Analysing each of these routes on the UK and European Union **("EU")** level enriches our understanding of how Ofcom's intervention could alleviate these issues from a regulatory perspective.

(a) Civil Claim against perpetrators for Online Hate

Due to the absence of legislation on online hate speech in the UK, data subjects have to resort to existing law that coincides with online hate speech. A prime example is defamation in tort law. Although defamation is not a form of hate speech, this civil misdemeanour is broad enough to include some comments which might be categorised as hate speech.

Defamation refers to statements (whether verbal or written) damaging the reputation of a person.[16] Since defamation encompasses a derogatory element that deeply wounds victims (as identified in the definition above), defamatory statements tend to express hate speech too. However, using this area of law to seek judicial redress for online hate is flawed in a number of ways.

[16] Health and Safety Executive, 'Defamation: libel and slander' (*HSE*, 2020) <https://www.hse.gov.uk/enforce/enforcementguide/court/reporting-defamation.htm> accessed 11 March 2021.

In defamation cases, there must be a direct link between the perpetrator and the victim. In other words, the defendant must be identifiable in order to raise a claim. This is an issue in the online sphere as the anonymity afforded by social media companies allows individuals to masquerade their real persona through fake accounts or fictional characters.

The real issue, however, lies in the elements that must be satisfied by the claimant in a defamation lawsuit, which make the threshold of accountability much more challenging to reach. This is exemplified in the recent case of *Miller v Turner*,[17] where the defendant published a series of tweets expressing hatred towards claimants by describing them as "Neo-Nazi anti-Semites" as well as "super fascists".[18] According to the definition of hate speech proposed by the Alan Turing Institute, the post expresses prejudice against both claimants, one of whom has a strong attachment to his Jewish identity, the other who is an avid supporter of minority racial groups. Nevertheless, in spite of the tweets' demonising statements, Collins Rice LJ was not satisfied that defamation took place. This is due to the language and context of the tweet bearing the semblance of an 'honest opinion' which exonerates the defendant of liability according to s.3 of the Defamation Act 2013.[19] However, should honest opinion, as hateful and derogatory as it is, be justified? It is argued this should not be the case where the gravitas of harm caused by the post far exceeds the importance of legitimate opinion. Harm in this instance refers to the emotional turmoil caused by the tweets not only on the claimants but on the online Jewish community as well.

On this basis, the paper views the proposed Online Safety Bill as a way to curtail online hate speech through a new conceptual apparatus akin to conduct crimes. The emphasis of tort law on the result (i.e., serious harm to the reputation of the claimant)[20] undermines the severity of the conduct itself (online hate speech), which undoubtedly has an effect on numerous victims. These victims are not necessarily identifiable, and would thus be prevented from pursuing tort law as an avenue for redress. Here, the need for alternative methods of accountability

[17] [2021] EWHC 2135 (QB).
[18] ibid, p. 20.
[19] Defamation Act 2013, s 3.
[20] Defamation Act 2013, s 1.

could not be more pronounced. Neither would have the claimants raised a defamation lawsuit if the hateful tweets were removed *ab initio* by Twitter. Both victims' true needs centre on merely deleting the post, however, the lack of legal frameworks curtailing online hate speech forced the claimants to resort to tort law. This is not withstanding the extortionate cost of bringing a civil lawsuit which, to the claimants' misfortune, did not result in a favourable outcome. This underscores that the Defamation Act is ineffective for online hate speech victims due to setting a high threshold of accountability. Additionally, the redress mechanisms (monetary remuneration) available to victims are not responsive towards their true needs where permanent reputational loss has been initiated.

(b) Civil claim against social media companies for online hate

Bringing a civil claim against social media companies for online hate speech as an accountability framework, has been shaped by EU law through the E-Commerce Directive (**"Directive"**).[21] Despite Her Majesty's government announcing that the E-Commerce Directive no longer applies to the UK following the end of the Brexit transition period, provisions on intermediary liability have been transposed into UK law.[22] Thus, discussing the Directive remains highly relevant to the UK.

Intermediary liability refers to holding social media companies accountable for illegal/legal yet harmful content emanating from users on their platforms. The rationale underpinning this liability framework is that publishers should be primarily responsible for perpetrators' hateful content since they provide the platform allowing such content to disseminate. There are two types of obligations imposed on intermediaries; reactive and proactive monitoring of content. Reactive monitoring refers to removing content after a post has been published.[23] The removal is a *reaction* to the post's illegal content which typically occurs via notice and take-down

[21] E-Commerce Directive (n 3).
[22] Department for Digital, Culture, Media and Sport, 'Guidance: The E-Commerce Directive and the UK' (*HM Government*, 18 January 2021) <https://www.gov.uk/guidance/the-ecommerce-directive-and-the-uk> accessed 22 January 2021.
[23] Kerie Kerstetter, 'Compare and Contrast: Proactive vs. Reactive Governance' (*Diligent Insights*, 29 October 2019) <https://insights.diligent.com/governance-intelligence/compare-contrast-proactive-vs-reactive-governance> accessed 9 October 2020.

obligations. The advantage of this approach is that it ensures careful scrutiny of material before deciding its illegality. The drawback, however, is that hate speech escalates at an exponential rate and take-down might thus be too late as the victim would have already faced severe damage. Proactive measures, on the other hand, refer to taking down content before being published thus preventing content to be seen by users navigating the platform. Whilst this prevents the dissemination of hate speech, over-moderation of materials can potentially ensue, thus infringing on freedom of expression.[24]

Provisions relating to intermediary liability in the E-Commerce Directive stipulate that host providers are not obliged to proactively monitor content by "actively seeking facts or circumstances indicating illegal activity"[25] pursuant to Article 15. However, upon being notified of illegal activity and not acting expeditiously to remove such illegal content, the host provider will be held liable according to Article 14(1).[26] Therefore, the knowledge standard imposed on social media companies in the E-Commerce directive is clear; there is no proactive obligation to monitor hateful content unless the intermediary service provider is notified, at which point liability arises if reactive measures are not deployed.

Despite the stance of the Directive being relatively clear, the interpretation of the Directive by the Court of Justice of the European Union **("CJEU")** in subsequent cases made the application of the Directive highly convoluted. The CJEU adjudicates on legal disputes submitted by one of its member states, often through a preliminary hearing. Such hearing is intended to clarify the stance of EU law as an overriding source of constitutional sovereignty across all EU member states.[27]

This is exemplified in the case of *Eva Glawischnig-Piesczek v Facebook Ireland Limited*,[28] where Mrs. Eva; the chair of a parliamentary party in Austria, sought an injunction obliging

[24] Chara Bakalis and Julia Hornle (n 13), p. 5.
[25] E-Commerce Directive (n 3), art. 15.
[26] ibid, art. 14(1).
[27] Raffaele Bifulco and Alessandro Nato, 'The Concept of Sovereignty in the EU – past, present and future' (*Reconnect - European Commission*, 30 April, 2020) <https://reconnect-europe.eu/wp-content/uploads/2020/05/D4.3.pdf> accessed 8 August 2021.
[28] Case C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] CJEU OJ C-104.

Facebook Ireland to remove a comment published on a social media network harmful to her reputation in addition to posts carrying identical and/or equivalent content.

The CJEU ruled that the Directive permits member states to order the removal of identical and equivalent information previously declared to be illegal, only if the equivalent content remains "essentially unchanged to the illegally declared substance" and contains the elements specified in the injunction.[29] The knowledge standard in this case therefore remains similar to the threshold stipulated in the Directive; SMC's do not have a general obligation to monitor illegal content, but can be obliged through injunctions to actively take down materials relating a specific post that has been declared illegal.

This knowledge standard has been significantly widened in the case of *Loreal v eBay* which created a new threshold referred to as 'constructive knowledge'.[30] This is different from 'actual knowledge' since, even if an intermediary service provider is not notified of an illegal post, the judgment indicates that there are certain instances when an intermediary service provider 'ought to have recognised' the illegality of the material, thus being obliged to initiate expeditious take-down.[31] Despite the case not relating to hate speech, it delineates the scope of article 14 of the Directive which applies to all intermediary service providers regardless of the factual backdrop and nuance of the specific case. The CJEU ruled that if a 'diligent economic operator' ought to have realised the illegality of the material and did not act expeditiously to take it down, the intermediary service provider will be held liable.[32] This contravenes the knowledge standard stipulated in the Directive that has been affirmed in *Eva* in relation to SMC's not being obliged to proactively monitor content. Van der Sloot justifies the judgment by stating that the CJEU is encouraging proactive monitoring of content by employing automated detection technologies.[33]

---

[29] ibid, Opinion of F. Biltgen J, para 39.
[30] Case C-324/09 *L'Oreal SA and Others v eBay International AG and Others* [2011] I-06011.
[31] ibid, para 105.
[32] *L'Oreal SA and Others v eBay International AG and Others* (n 29), paras 120 – 124.
[33] Bart Van Der Sloot, 'Welcome to the Jungle: The Liability of Internet Intermediaries for Privacy Violations in Europe' 6 [2015] JIPITEC, para., 26 < https://www.jipitec.eu/issues/jipitec-6-3-2015/4318> accessed 13 March 2021.

Therefore, it is concluded that the *prima facie* knowledge standard stipulated in the Directive is clear, although it has been significantly widened in *Loreal v eBay* in an apparent attempt by the CJEU to increase the liability threshold placed on intermediary service providers.

It is also essential to note that the knowledge standard stipulated in the E-Commerce Directive is the standard that shall apply to the UK. This is been expressly stated in the Online Harms White Paper,[34] and thus calls for clarification from the government on whether CJEU judgments will be taken into account by Ofcom when evaluating the knowledge standard expected of SMC's for taking down posts.

(c) Issues arising from the imposition of a new Knowledge Standard

Through the imposition of a more stringent knowledge standard in the CJEU judgment of *Loreal v Ebay*,[35] it is foreseeable that social media companies will have recourse to Artificial Intelligence (**"AI"**) systems to proactively monitor content. AI refers to an automated computer-based system capable of performing tasks requiring human intelligence, which in this context would replace a physical human verifier when it comes to monitoring content.[36] This is highlighted in the case of *Eva* where the CJEU emphasizes that having recourse to AI would not require host providers to conduct an independent assessment of illegal content.[37]

However, having exclusive recourse to AI systems can create adverse effects. These systems can potentially create disproportionate take-down of posts as automated tools can swiftly indicate identical content, but cannot indicate equivalent content phrased in an indirect way, especially if phrased in a context-sensitive way as a form of political satire, parody or irony.[38] Thus, inevitably, this element must involve an aspect of human review to determine illegality,

---

[34] HM Government, 'Guidance: The eCommerce Directive and the UK' (n 22), grey box 1.
[35] *L'Oreal SA and Others v eBay International AG and Others* (n 30).
[36] Darrell M. West, 'What is Artificial Intelligence' (*Brookings*, 4 October 2018) <https://www.brookings.edu/research/what-is-artificial-intelligence/> accessed 11 March 2021.
[37] *Eva Glawischnig-Piesczek v Facebook Ireland* (n 28), para 46.
[38] Michael Herz & Peter Molnar (eds), *The Content and Context of Hate Speech: Rethinking Regulation and Reponses* (Cambridge University Press 2012).

as resorting solely to automated technologies can curb the right to freedom of expression in a platform as ubiquitous as Facebook.

IV. How would Ofcom's intervention change the liability regime?

If Ofcom is appointed as the regulator for online harms as has been proposed in the Online Safety Bill,[39] victims would not need to sue perpetrators directly for hate speech (via tort law) nor would they need sue social media companies for failing to take down posts expeditiously (as compliant with the E-Commerce Directive). Rather, they would be able to rely on a central regulator who has been exclusively delegated with the task of overseeing the activities of intermediary service providers and can hold them accountable for not complying with predetermined codes of conduct. The potential appointment of Ofcom would therefore not only ameliorate the inadequacies of the existing routes for judicial redress, it would create a regime that is responsive to victims' true needs. These needs are centred on the expeditious removal of illegal content without incurring extortionate costs through litigation. Ofcom as a telecommunications regulator would be able to fulfil this by holding social media companies accountable to higher standards as concomitant with the scale, severity and complexity of the situation.[40] According to Singh and Mara, holding a body 'accountable' is different from holding a body 'liable': the former refers to a broader social responsibility, whilst the latter refers to a narrowly construed legal responsibility.[41] This resonates with Ofcom's social mission of combatting harms inflicted on society.[42]

More importantly, the role of Ofcom as a central regulator for policing online harms becomes akin to the role of the Crown Prosecution Service which polices harmful behaviour in society. Therefore, mere 'bystanders' who are not directly affected by hateful content would

[39] Draft Online Safety Bill (n 7), part 4.
[40] Online Harms White Paper (n 5), box preceding para 2.54.
[41] Vikram Singh and Prashant Mara, 'Liable v Accountable: How Criminal Use of Online Platforms and Social Media poses challenges to intermediary protection in India' (*BTG Legal*, 5 May 2020) <https://www.mondaq.com/india/social-media/928106/liable-vs-accountable-how-criminal-use-of-online-platforms-and-social-media-poses-challenges-to-intermediary-protection-in-india> accessed 3 January 2021.
[42] Online Harms White Paper (n 5), box 5.

potentially be able to raise a claim to Ofcom. This underscores the role of Ofcom in not only ameliorating harm at an individual level, but also at a wider societal level by targeting generalised comments and radicalised hate movements – as emphasized by Bakalis.[43]

However, with the potential introduction of a regulator, a range of implications arise. Firstly, Frosio questions whether shifting blameworthiness from perpetrators of hate speech to the platforms allowing such hate speech to penetrate neglects the 'root cause' for disseminating vitriolic content; perpetrators.[44] This paper disagrees with Frosio's stance as seeking criminal liability does not appeal to victims' true need of initiating expeditious take-down to prevent further emotional harm rather than punishing the perpetrator. Hence, by imposing a higher threshold of accountability on social media companies through a suite of enforcement measures such as issuing fines,[45] expeditious take-down of illegal posts is initiated. This in turn, *likely* reduces the need for seeking criminal liability.

Nevertheless, the distinction between illegal/criminal content and legal yet harmful content is a key contention that could arise with Ofcom's potential appointment. Woods emphasizes on how using both categories as a "proxy for identifying harm" is a problem as criminal law is not designed to *measure* harm but rather decides what society wants to penalise.[46] This is exacerbated with the Online Safety Bill's vague definition of the 'legal yet harmful category' which is invoked if the provider has "reasonable grounds to believe that the nature of content [poses] a significant adverse physical or psychological impact on a child/adult".[47] The lack of clarity in what could be ostensibly classified as 'adverse' gives Ofcom unfettered discretion when evaluating harmfulness. This can potentially result in over-moderation, thus infringing on freedom of expression.

---

[43] Chara Bakalis, 'Rethinking cyberhate laws' (n 4), p. 3.
[44] ibid., page 7, para 3.
[45] ibid., para 4.43.
[46] SCL Student Bytes, 'Making sense of tech law podcast - Episode 7: Reducing Online Harms: A Statutory Duty of Care (Lorna Woods, 9:32 min.) <https://podcasts.apple.com/ni/podcast/episode-7-reducing-online-harms-a-statutory-duty-of-care/id1550739388?i=1000516666382&l=en> accessed 12 August 2021.
[47] Draft Online Safety Bill (n 7), s 45(3).

V. How effective is the Duty of Care?

The White Paper proposals envisage a duty of care on social media companies to tackle online harms effectively and initiate more proportionate and expeditious take-down[48]. This has also been proposed in the Online Safety Bill where the Secretary of State has been delegated with the power to create secondary legislation by incorporating Ofcom's codes of practice.[49] This section does not evaluate the substantive content of the interim codes of conduct issued by Ofcom,[50] as they have not been finalised yet and are thus not legally enforceable. Rather, by critiquing the effectiveness of the duty of care as a vehicle for enhancing intermediary accountability, an understanding is gauged as to the duty's efficacy in light of its presumed aim.

The effectiveness of the duty is ostensibly seen in enlarging the responsibility placed on social media companies who must not merely adhere to codes of conduct (that flesh the duty), but also assume a wider responsibility for tackling online harms. The duty carries social connotations which are beneficial and instigate a higher degree of responsibility, but are equally nebulous, vague and broad. To evaluate the efficacy of the duty of care, an essential question must be answered: will the duty compel social media companies to react more proactively when combatting online hate speech?

To answer this question, the purpose of tort law which created the 'duty of care' must be deliberated. According to Morrow, the law of tort was created as a middle ground between contractual and criminal liability.[51] Certain acts in society that ought to be rebuked could not have been regulated through contract or criminal law due to falling outside the remit of both. Consequently, tort law was created to fill this gap and ensure a new mechanism that created liability for what were once 'elusive' notions such as 'nuisance' and 'trespass to land'.

---

[48] Online Harms White Paper (n 5), para. 4.4.
[49] Draft Online Safety Bill (n 7), s 29(1).
[50] HM Government, 'Guidance: The eCommerce Directive and the UK' (n 22), para 2.46.
[51] Karen Morrow, 'Tort and Regulatory Law in England and Wales' [2018] 19 Tort and Insurance Law (TIL),<https://link.springer.com/chapter/10.1007/978-3-211-31134-9_5> accessed 14 March 2021.

When assessing whether the duty would sufficiently compel social media companies to react differently, the benefit of the duty is seen through holding social media companies liable not only for a positive act (e.g., disproportionate take-down of posts), but also for omissions (e.g., failing to take down harmful posts). However, how wide is too wide? A plethora of case law following the seminal case of *Donoghue v Stevenson*,[52] show that the scope of the 'duty' has been delineated incrementally, albeit being often widened or restricted according to the factual backdrop of the case. Even after the scope of the duty has been demarcated through a barrage of cases succeeding *Caparo* in 1990,[53] Lord Toulson explains in *Michael v Chief Constable of South Wales Police*[54] which took place in 2015 that "the concepts of proximity and fairness are not susceptible to any definitions which would make them useful as practical tests".[55] This shows how the duty should be defined more narrowly as it creates an extra layer of responsibility on top of following the codes of conduct. Had the codes only been issued without the overarching duty, this would not have been a concern given the clarity and specificity of the codes which do not have a "wide margin of appreciation".[56] Van der Sloot further questions the efficacy of the duty on a practical level, given that Ofcom adopts a 'comply and explain' procedure when abiding by the codes.[57] The lack of strict liability therein raises the question of whether breaching the code and having an explanation would also absolve the company in question of liability for the duty.

VI. Is there any alternative? Liability through Contract Law

With the flaws of the tort law model, an alternative framework for curtailing online hate speech can be achieved through contract law.

Four elements are needed for a valid contract: offer, acceptance, consideration and intention to create legal relations. Consideration is not at issue as case law shows that this element is

[52] [1932] UKHL 100; [1932] AC 562.
[53] *Caparo v Dickman* [1990] 2 AC 605 (HOL).
[54] [2015] UKSC 2
[55] ibid [106] (Lord Toulson).
[56] Bart Van Der Sloot, (n 33), para 99.
[57] Online Harms White Paper (n 5), 3.22.

not satisfied only through monetary means. According to *Currie v Misa*,[58] "a valuable consideration, in the sense of the law, may consist either of some right, interest, profit, or benefit accruing to the one party, or some forbearance, detriment, loss or responsibility, given, suffered or undertaken by the other". In the context of social media companies, the benefit provided by the user would be his/her personal data. This is construed as a form of "profit" for social media companies as the data is sold to third parties and advertising agencies. The detriment faced by the user is that he sacrifices his privacy. This argument has been affirmed by Zuboff who, in her seminal book 'The Age of Surveillance Capitalism',[59] unravels the competitive dynamics of the data market. She concedes to social media companies acquiring "ever-more predictive sources of behavioural surplus" to "nudge, coax, tune and herd behaviour toward profitable outcomes".[60] Zuboff's work accentuates on the notion of 'mutuality' in consideration which seems to be undermined when considering the exploitative practices social media companies adopt through mass surveillance.[61]

In light of this, the benefits of the proposed framework are evident in the fact that it is concomitant with the business model of social media companies.[62] Facebook, Twitter and TikTok amongst others sell personal data without any 'exchange' or consideration. Harvesting data has been at the core of the ostensible 'free services' users receive, but data is the valuable asset allowing these companies to prosper and thrive.[63] This lack of mutuality is addressed through the proposed framework as both parties' interests are protected: social media companies get the data they want, whilst users have a legal mechanism that protects their interest of being protected from hate speech. This would be achieved through the terms and conditions stipulating when the content is construed as hateful, thus triggering breach of contractual rules in instances of hate speech which mandates instantaneous take-down of posts. ARTICLE 19, a human rights organisation, is an avid supporter for side-stepping regulation through contract. In their comprehensive policy brief, they shed light on how

---

[58] (1875) LR 10 Ex 153 (HOL).
[59] Shoshana Zuboff, *The Age of Surveillance Capitalism* (1st edn, Profile Books Ltd, 2018).
[60] ibid., ch 4, page 76.
[61] ibid., ch 4, page 83.
[62] Roel Wieringa, 'A Business Model of the Facebook ecosystem' (*The Value Engineers*, 5 June 2020) <https://www.thevalueengineers.nl/a-business-model-of-the-facebook-ecosystem/> accessed 3 October 2020.
[63] Joris Toonders, 'Data is the new oil of the digital economy' (*Wired*, 6 September 2020) <https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/> accessed 11 March 2021.

regulation by contract eliminates state intervention through disproportionate censorship.[64] The proposed framework is hence not only more consistent with international standards on freedom of expression, but it puts an end to the State or public organisations being commissioned by the State (e.g., Ofcom) to be the "ultimate arbiter of what constitutes permissible expression". This eliminates the power asymmetry between companies and the state, thus enshrining users' freedom of expression by preventing over-moderation and onerous codes of practice, thus protecting minority viewpoints from a human rights lens.[65]

The second benefit this framework presents is that it reduces the need for community guidelines. Numerous instances relating to not taking down a post in spite of it being unlawful have arisen due to the legal regime squarely contradicting the business model (i.e., community guidelines).[66] This issue is addressed through contract law in a number of ways. Firstly, the role of Ofcom can be re-orchestrated by issuing template contracts that *must* be issued by social media companies to their users. The provided template is inalienable (cannot be changed), but can be built upon by voluntarily adding more terms to the provided template. Therefore, the initial template protects the legal interests of both parties pertaining to proportionate take-down and hateful content, whilst the added terms and conditions allow social media companies to incorporate their community standards, which become not mere guidelines, but legally enforceable contractual terms. This ingenious mechanism is the ideal way to reconcile legal interests with business interests. It is imperative to note that the added terms and conditions must not contravene the *prima facie* legal position stipulated in the model terms. This is an element that can be overseen by Ofcom to ensure regulatory compliance.

[64] 'Policy Brief - Side-stepping rights: Regulating Speech by contract' (*Article 19*, 2018) <https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB.pdf> accessed 9 August 2021.
[65] ibid., p. 35.
[66] Kenny Novak, 'Why Facebook Deletes Yours Posts for Community Standards' (*Boostlikes*, 13 July 2019) <https://boostlikes.com/blog/2019/07/deletes-post-community-standards> accessed 7 March 2021.

Furthermore, the practice of issuing model contracts that can be built upon has been prevalent in the field of data protection and is referred to as 'privacy by design'[67] – due to protecting legal interests by default through designing the contractual rules. Such contracts have been issued by the Information Commissioner's office for data exports outside the European Economic Area to protect the data privacy, integrity and confidentiality of users in the European Union.[68] In this regard, replicating this practice to online hate speech could be a promising step towards addressing the issues associated with the tort law model.

The drawbacks of the contractual mechanism lie in contracts often being construed very strictly due to courts adopting a literal interpretation when assessing contractual duties.[69] While implied terms can be attached to contracts in certain instances, this is very limited in case law. The need for mentioning everything explicitly might limit the ability of victims to invoke their interests and might disregard the sheer volatility and context-sensitivity of hate speech which can carry a wide range of different connotations. Therefore, Ofcom as a regulator, must precisely define what constitutes 'hate speech' and how the right to freedom of expression can be delineated contractually.

Ultimately, the benefit of the contract law route is seen in the fact that it would enhance Ofcom's role as the watchdog, as despite the contract operating *ipso facto*, compliance would be overseen by Ofcom and damages would be similarly issued for contractual breach. These monetary/non-monetary damages will be decided based on the gravitas of harm which contravenes a 'baseline threshold' that is delineated contractually and must not be exceeded. Terms and conditions would also have much more substance and clarity in contrast to the 'duty of care' which is elusive and might be 'all encompassing' depending on the unilateral discretion of Ofcom which creates lack of certainty for social media companies.

---

[67] European Commission, 'Standard Contractual Clauses' (*European Commission*, 21 September 2020) <https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/standard-contractual-clauses-scc_en> accessed 30 September 2020.

[68] Information Commissioners Office, 'Build a Controller-to-Controller Contract' (*ICO*, 2020) <https://ico.org.uk/for-organisations/dp-at-the-end-of-the-transition-period/build-a-controller-to-controller-contract/> accessed 24 October 2020.

[69] David Capps, 'Interpretation of Contracts under English Law' (*Ashurst*, 18 August 2020) <https://www.ashurst.com/en/news-and-insights/legal-updates/interpretation-of-contracts-under-english-law/> accessed 13 January 2020.

VII. Intersection between Tort Law and Contract Law

While both tortious and contractual mechanisms seem mutually exclusive due to being invoked in different circumstances, this paper argues that both routes should be used in tandem to curtail online harms. As such, both frameworks should be available and invoked depending on the context of hate speech.

This paper recommends, as aforementioned, that the codes of conduct be embedded contractually to trigger automatic breach in instances of lack of compliance. This recommendation is subject to two limitations. Firstly, the White Paper adopts a 'comply or explain' procedure which changes through this mechanism, as not complying with contract initiates a breach which is not subject to justification. The second limitation emanates from the question: if the current approach already issues fines for lack of compliance, what difference will contract law create? The answer is ostensibly seen from the perspective of victims, as if they think that their rights have not been sufficiently enforced by the regulator, they can use contract law to directly raise a claim against social media companies. While Ofcom is intended to remove the onus from users and eliminate the costs associated with private litigation, resorting to this pathway is foreseeable given that Ofcom only deals with "systemic issues causing harm".[70]

Tort law would create an additional layer for enforcing elements that cannot be imposed contractually through the duty of care. Thus, elements not falling within contract law, can be invoked through the tortious mechanism, albeit with clearer guidelines on when a duty of care arises.

VIII. Conclusion

This paper concludes that the new regulatory regime proposed in the White Paper addresses many of the issues faced in the current liability framework. Through a central regulator, the onus is removed from victims of hate speech who do not need to sue perpetrators (through

[70] Online Harms White Paper (n 5), para 4.37.

tort law) or social media companies (through the E-Commerce Directive). The benefit of the regulator is underscored in targeting generalised forms of hate speech and radicalised movements, which extends the responsibility of the regulator through targeting wider forms of hate speech that damage the social fabric of society. Through the duty of care scheme which governs the relationship between social media companies and victims, mutual rights are enforced more effectively and a layer of social responsibility is attached to SMC's.

However, key issues arise with the current liability framework. Such issues encompass the enormity of the duty which can potentially be all-encompassing, thus creating uncertainty for social media companies on when the duty is triggered. To ameliorate this issue, the paper recommends resorting to contract law for 'baseline' or 'clear-cut' instances of hate speech through model contracts which can be built upon to reconcile legal and business interests. More context-sensitive forms of hate speech would subsequently trigger the duty of care, albeit the scope of the duty has to be more narrowly defined to compel social media companies, *prima facie*, to act more proactively.