EJLT European Journal of Law and Technology

Differentiating between Harm to Users and Third Parties in the UK's Online Safety Regulations

Ellie Colegate[*]

Abstract

Since 2022, there has been a notable increase in individuals taking part in offline disturbances after being influenced by user-generated content on social media. While it has previously been shown that online content can harm individuals, the consideration of harm is typically confined to the singular viewing user. However, the increase in offline demonstrations has shown the potential for harm to third parties, a consequence often overlooked in regulatory initiatives. The UK's Online Safety Act 2023 seeks to curb content that could cause harm to users by mandating actions platforms must take. The rise in adverse impacts of content that could harm third parties raises questions about the how suitable the Act's provisions are when harm or risk of harm – the reduction of which is at the core of the regulations – reaches beyond a viewing user and generates offline consequences to third parties and property. Centring on the lack of consideration for third-party harms as an outcome of user-generated online content, this paper explores how and where these occur, identifies the deficits present in current regulations, and offers pragmatic ways forward so a balance between viewing users and third parties can be achieved.

Keywords: content regulation, Online Safety Act 2023, online harms, viral content, harm to online users, third-party harms.

[*] PhD Candidate at the School of Law and the EPSRC Horizon Centre of Doctoral Training, University of Nottingham. This work was supported by the Engineering and Physical Sciences Research Council Grant No. EP/S023305/1. I would like to thank the anonymous reviewers for their insightful comments, which greatly enhanced this work and clarified my ideas. I also want to thank Professor Richard Hyde for his feedback and support. Any errors are my own.

1. Introduction

Discussions and commentary concerning online content generally centre around the impact it can have on users consuming and interacting with it, while regulations and guidance are typically tailored towards the impact a piece of content could have on a user exposed to it. Thus, platforms now decide whether to remove a piece of content from its services based on these descriptions and thresholds.[1] However, in recent years, there has been an increasing trend of third-party individuals being harmed due to content circulating online,[2] a new phenomenon that, arguably, was not foreseen within regulatory efforts.

Various instances of activism enabled or assisted by online content have been seen over the last decade.[3] Online calls to action and hosted content resulting in offline demonstrations, protests and actions – positive and negative – are not new consequences of user-generated online content.[4] However, there has been a notable increase in the intensification of offline demonstrations as a direct consequence of content seen on social media platforms, with users reporting that if it were not for platforms encouraging them, they would not have engaged in actions that cause third parties harm. In parallel to such events, the development of regulations and laws to ensure platforms control what they host online has also been taking place. The UK's Online Safety Act 2023 ('the Act') mandates that platforms must remove content from public view should it pose a risk of harm to users, and provides an accompanying legislation-specific definition of 'harm' to guide regulatory decisions. With instances of offline third-party harm occurring despite the introduction of provisions, and debates being had around the extent to which new laws will stifle legitimate protest,[5] it is essential to determine whether third-party harms should be acknowledged as part of the regime overall.

---

[1] Online Safety Act 2023.

[2] 'TikTok Frenzies "Putting Police and Schools under Strain"' BBC News (23 September 2023) <https://www.bbc.com/news/technology-66888029> accessed 27 September 2023; Sammy Gecsoyler, 'Police Leader Calls on TikTok to Investigate Oxford Street "Robbery" Campaign' *The Guardian* (15 August 2023) <https://www.theguardian.com/uk-news/2023/aug/15/police-leader-calls-tiktok-investigate-oxford-street-robbery-campaign> accessed 29 November 2023; Holly Evans, 'TikTok Disorder Spreads to Southend as Teens in Balaclavas Stopped by Police' *The Independent* (11 August 2023) <https://www.independent.co.uk/news/uk/crime/tiktok-riot-southend-oxford-street-b2391405.html> accessed 29 November 2023.

[3] Howard Phillip et al, 'Opening Closed Regimes: What Was Role of Social Media During the Arab Spring?' (Project on Information Technology & Political Islam, 2013) Working Paper; Jessi Hempel, 'Social Media Made the Arab Spring, But Couldn't Save It' Wired.com (26 January 2016) <https://www.wired.com/2016/01/social-media-made-the-arab-spring-but-couldnt-save-it> accessed 7 February 2020.

[4] Monica Anderson et al, 'Activism in the Social Media Age' (Pew Research Centre, 2018); Kristine Mitchell, 'Digital and Online Activism' (May 2020). <https://en.reset.org/knowledge/digital-and-online-activism> accessed 27 January 2020.

[5] Mark Leiser and Edina Harbinja, 'Why the UK Proposal For a "Package of Platform Safety Measures" Will Harm Free Speech' <https://techreg.org/index.php/techreg/article/view/53> accessed 18 August 2021.

To satisfy their new obligations and reduce the likelihood of the content they host online having adverse impacts, it is vital that platforms understand to whom a risk of harm is presented: a viewing user or a third party. This paper utilises the Act's provisions and significant documented cases of third-party harm to elucidate the different types of harm that can extend beyond viewing users. It will investigate the contrasting relationship between harm inflicted on a viewing user and harm to a third party, emphasising its importance in effectively mitigating harm to individuals through new regulations across various domains.

Firstly, this paper will explore and showcase the understanding of 'harm' within the UK framework to establish the current consideration for the third parties present. The difference between harm to viewing users and third parties will then be explored as the primary reference point for this paper's discussions and recommendations. As part of this, three types of harm to third parties will be showcased. The discussion will then evaluate regulatory deficits, examine the significance of underlying platform architecture, and highlight the need to balance the protection of viewers with the interests of third parties. Recommendations for how third parties can be acknowledged in the regulatory framework, concluding remarks and signposting for areas of further work will be set out thereafter.

2. Understanding Harm in the UK Regulatory Environment – the Online Safety Act 2023

Introduced in October 2023, the Online Safety Act has been promoted as the regulatory framework to make the UK the 'safest place in the world to be online'.[6] At just over 300 pages, and containing 241 sections and 17 Schedules, the aim of the Act is to comprehensively regulate online platforms and services.[7]

2.1 Section 234

The concept of harm sits at the heart of the provisions,[8] with the harmful impacts and adverse effects of exposure to online content driving the work of parliamentarians since the 2019 Online Harms White Paper,[9] which introduced the idea of the 'digital duties of care'. The Act mandates that user-to-user services[10] must remove content that poses a risk of harm to its users.[11] Moving away from existing regulations in the

[6] Department for Science, Innovation and Technology et al, 'UK Children and Adults to Be Safer Online as World-Leading Bill Becomes Law' GOV.UK <https://www.gov.uk/government/news/uk-children-and-adults-to-be-safer-online-as-world-leading-bill-becomes-law> accessed 26 October 2023.
[7] Ofcom, 'Online Nation Report 2022' (2022) <https://www.ofcom.org.uk/__data/assets/pdf_file/0023/238361/online-nation-2022-report.pdf>.
[8] Online Safety Act, s 234.
[9] The Department for Digital, Culture, Media and Sport, *The Online Harms White Paper* (2019).
[10] Online Safety Act, s 3.
[11] ibid 7.

broader legal area, the Act provides a conceptualisation and understanding of what constitutes harm and potentially harmful content for both adult and child users, encapsulating the concept of 'online harms'. Section 234 outlines the legislative and guiding definition of 'harm', providing platforms with a benchmark against which to recognise and categorise content as harmful or not. The definition provided within the interpretation of the Act succinctly states that harm can be 'physical or psychological'[12] in nature, and indicates that for content to be considered harmful overall, the risk or likelihood of either type occurring as a consequence of interaction would need to be presented.

3. Recognising Harms to Third Parties and where these Occur

In order to understand why regulations might need to allow for and recognise the impacts that online content can cause third parties, there needs to be an examination of how these differ from the impacts already considered in relation to viewing users and how they can arise. As indicated, harms that have been experienced by individual viewing users are those that can be seen to drive the works of governments and regulatory initiatives internationally as specific online safety regulations are introduced and enacted.[13] This section now explores the differences between harm to viewing users and harms to third parties, using examples of recent situations where third-party harms can be categorised into recognisable types.

Offline demonstrations and mobilisations of users have been increasingly reported as the reach of the internet has grown. Whilst the most notable mobilisations of individuals following the sharing of user-generated content have been widespread in their form and reach, such as the Black Lives Matter demonstrations in 2020[14] and protests for environmental protection,[15] these can vary in scale and severity.[16] Online environments and platforms can serve as both planning hubs for demonstrations[17] and a source of inspiration for users to mobilise as user-generated content both contains 'calls to action' and normalises, shares and promotes offline responses.[18]

---

[12] ibid s 234(2).
[13] Department for Science, Innovation and Technology et al (n 6).
[14] Guobin Yang, 'Narrative Agency in Hashtag Activism: The Case of #BlackLivesMatter' (2016) 4 *Media and Communication* 13.
[15] Niels G Mede and Ralph Schroeder, 'The "Greta Effect" on Social Media: A Systematic Review of Research on Thunberg's Impact on Digital Climate Change Communication' (2024) 18 *Environmental Communication* 801.
[16] John D Gallacher, Marc W Heerdink and Miles Hewstone, 'Online Engagement Between Opposing Political Protest Groups via Social Media Is Linked to Physical Violence of Offline Encounters' (2021) 7 *Social Media + Society* 205630512098444 5.
[17] Paul Gill et al, 'Terrorist Use of the Internet by the Numbers: Quantifying Behaviors, Patterns, and Processes' (2017) 16 *Criminology & Public Policy* 99; Laura GE Smith et al, 'Digital Traces of Offline Mobilization' (2023) 125 *Journal of Personality and Social Psychology* 496.
[18] Hedy Greijdanus et al, 'The Psychology of Online Activism and Social Movements: Relations between Online and Offline Collective Action' (2020) 35 *Current Opinion in Psychology* 49; Laura

Whilst much of the content that contributed to the examples used in this paper were viral in nature – thus attracting media attention when offline demonstrations caused significant impacts – smaller digital traces can also be key to mobilisations,[19] with polarising content and echo chambers[20] often inspiring users to act in a collective way.[21]

Across all of the examples explored here, social media and user-generated content hosted therein has had a significant role in mobilising individuals and the third-party harms that subsequently arose. Between February and March 2023, online content that encouraged students in UK schools and colleges to participate in demonstrations of rebellion was shared on both TikTok and Facebook,[22] amplifying messages of protest and calls for action. A month earlier, in January 2023, mass gatherings and disturbances were mobilised by similar call-to-action content after the disappearance of a local woman in St Michael's on Wyre, Lancashire. These events, which may not have occurred without the encouragement provided by online material, led to third-party harms: a significant volume of online content was created and shared, which, in turn, inspired individuals to participate in unsanctioned investigations and engage in harassing behaviours they would otherwise have avoided. Likewise, looting and disorder on Oxford Street in London in August 2023[23] were prompted by online content instructing young people to take action on a specific date and time. In addition, content that both polarised users and inspired UK-wide violent offline

GE Smith et al, 'Digital Traces of Offline Mobilization.' (2023) 125 *Journal of Personality and Social Psychology* 496.

[19] Smith et al (ibid).

[20] Sarita Yardi and Danah Boyd, 'Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter' (2010) 30 *Bulletin of Science, Technology & Society* 316.

[21] At this point, it should be noted that malicious actors can drive online content to achieve specific goals, such as inciting offline demonstrations and disturbances. While this is feasible, it is beyond the scope of this paper and merits further exploration, given that the focus here is on third parties as affected actors, rather than on malevolent actors instigating demonstrations.

[22] Tom Sanders, 'Children "riot" in schools across the country to protest toilet rules' *Metro* (24 February 2023) <https://metro.co.uk/2023/02/24/pupils-riot-at-schools-across-the-country-during-toilet-rule-protest-18343826> accessed 6 December 2023; Abbie Wightwick, '34 Pupils Suspended after Unisex School Toilets Protest Shared on Tiktok' *Wales Online* (7 March 2023) <https://www.walesonline.co.uk/news/education/34-pupils-suspended-after-unisex-26405216> accessed 6 December 2023; Kevin Shoesmith, 'Hull's Kingswood Academy Damaged during Pupil "Disturbance"' BBC News (28 February 2023) <https://www.bbc.com/news/uk-england-humber-64792865> accessed 6 December 2023; Eleanor Busby, 'TikTok Challenged as School Pupils Plan More Copycat Protests' *The Independent* (6 March 2023) <https://www.independent.co.uk/news/uk/politics/tiktok-school-protests-toilets-uniform-b2295151.html> accessed 6 December 2023.

[23] Ben Ashton and Sam Corbishley, 'Chaos on Oxford Street as Police Clash with Teens amid TikTok Crime Spree Threat' *Metro* (9 August 2023) <https://metro.co.uk/2023/08/09/police-ramp-up-oxford-street-patrols-after-tiktok-crime-spree-threat-19306612> accessed 15 December 2023; Holly Evans, 'How a TikTok Craze Led to Hours of Chaos on London's Busiest Shopping Street' *The Independent* (11 August 2023). <https://www.independent.co.uk/news/uk/crime/oxford-circus-jd-robbery-rampage-tiktok-b2391303.html> accessed 10 December 2023.

demonstrations was seen in August 2024, with anti-immigration protests, unrest and riots occurring across the UK following misinformation on social media.[24] Most recently, influencers have encouraged attendance at demonstrations against changes to tax policies for farmers since November 2024, and likely increased the number of participants.[25]

Of all the examples explored, the connection to social media platforms and user-generated content was most overt in the case of St Michael's on Wyre. Here, the College for Policing retrospectively noted the specific use of TikTok to spread content related to the woman's disappearance, with related videos on TikTok gaining 270 million views in total.[26] The intense popularity of this content encouraged people to visit the town and undertake their own independent investigations, and this was a primary factor in the consequential social harm to third parties, with individuals being incorrectly accused of malicious activities and properties being trespassed upon, reaffirming the role of platforms as a controlling entity. The unprecedented attention given to the case highlighted that there is the potential for content relating to

[24] Amy-Clare Martin, Holly Bancroft and David Maddox, 'Nation Braces for Weekend of Far-Right Violence with 35 Protests in Wake of Southport Stabbing' *The Independent* (2 August 2024) <https://www.independent.co.uk/news/uk/crime/southport-far-right-protests-police-b2590232.html> accessed 9 September 2024; Marianna Spring, 'Did Social Media Fan the Flames of Riot in Southport?' BBC News (31 July 2024) <https://www.bbc.co.uk/news/articles/cd1e8d7llg9o> accessed 9 September 2024; Hannah Al-Othman et al, 'Five Arrested after 53 Police Officers Injured in Southport Riots' *The Guardian* (31 July 2024) <https://www.theguardian.com/uk-news/article/2024/jul/31/three-arrested-over-southport-riots#:~:text=More%20than%2050%20police%20officers,attack%20that%20killed%20three%20children.> accessed 9 September 2024; PA Reporters, 'Southport Residents Repair Wrecked Streets and Say Killed Girls 'didn't Deserve' Riot' *London Evening Standard* (31 July 2024) <https://www.standard.co.uk/news/uk/southport-people-humza-yousaf-merseyside-merseyside-police-b1173958.html> accessed 9 September 2024; Ben Quinn and Dan Milmo, 'How TikTok Bots and AI Have Powered a Resurgence in UK Far-Right Violence' *The Guardian* (2 August 2024) <https://www.theguardian.com/politics/article/2024/aug/02/how-tiktok-bots-and-ai-have-powered-a-resurgence-in-uk-far-right-violence> accessed 9 September 2024; Andrew Griffin, 'How a Few Twitter Posts on Elon Musk's X Helped Fan the Flames of Unrest and Rioting across the UK' *The Independent* (6 August 2024) <https://www.independent.co.uk/tech/elon-musk-uk-riots-southport-twitter-x-b2591725.html> accessed 9 September 2024.

[25] Jenny Kumah et al, 'Farmers March in Inheritance Tax Protest in London' BBC News (19 November 2024) <https://www.bbc.com/news/articles/czj71zyy934o> accessed 3 February 2025; James Tapper, '"We've Become the Voice of Agriculture": The Social Media Influencers Driving the Big Farming Protests' *The Observer* (17 November 2024) <https://www.theguardian.com/environment/2024/nov/17/weve-become-the-voice-of-agriculture-the-social-media-influencers-driving-the-big-farming-protests> accessed 3 February 2025.

[26] College of Policing, 'Independent External Review of Lancashire Constabulary's Operational Response to Reported Missing Person Nicola Bulley' (2023) 5 <https://assets.college.police.uk/s3fs-public/2023-11/Nicola-Bulley-independent-external-review.pdf>.

unfolding events to be promoted on platforms over other topics. The College indicated that content such as that which caused significant harm to the area and individuals may occur in future due to the behind-the-scenes promotion and recommendations by the platforms.[27]

3.1 Harms to Viewing Users

As shown above, the Act considers harm foremost as a phenomenon that affects viewing users rather than third parties. This fails to consider the impact upon the third parties, and leaves such impacts unregulated, which does not uphold the aim to reduce the various harms that can stem from online content.[28] To understand this further, the difference between third parties and viewing users, and the harms caused to both, need to be unpacked. Following the analysis of key recent examples, this paper categorises their potential adverse impacts into three types of harm: economic; physical; and social.

Harm to viewing users (but not economic harm) is already expressed and allowed for within the Act. This means that, for this paper, the viewing user is the individual interacting with content in the first instance, potentially acting on what they consume, and those who are at the forefront of considerations for platforms and regulatory stakeholders. These are accounted for by the legislation obligating platforms to undertake both adult[29] and child[30] risk assessments, utilise methods to verify the ages of those interacting on their services,[31] and have general regard for any potential adverse impacts or harms that their hosting and promotion of user-generated content could cause.[32] Potential harms to this group are legislated for under section 234, which provides a reference point for services when carrying out their duties and efforts to adhere to obligations.[33] In comparison, third-party harms are less extensively documented across literature and reports. However, as seen in the discussions above, the examples of harm to third parties are comparable to the former, yet not explicitly addressed in the Act.

When an individual viewing user experiences harm as a consequence of what they see or interact with online, it is typically an isolated reaction with no other parties needing to be involved for the individual to consider themselves harmed. For example, self-harm by a viewing user has repeatedly been recognised as a harm type that occurs due to content existing online.[34] When this occurs, it is an incident isolated

[27] *The TikTok Effect* (BBC Current Affairs, 2023) <https://www.bbc.co.uk/iplayer/episode/m001qp28/the-tiktok-effect> accessed 2 November 2023.
[28] Department for Science, Innovation and Technology et al (n 6).
[29] Online Safety Act, s 9.
[30] ibid s 11.
[31] ibid s 64.
[32] ibid s 7.
[33] ibid s 234.
[34] ibid s 62(3)–(4).

to the viewing user, with no immediate third-party, secondary harm occurring. This potentially arises in connection to a viewing user's family or friends, should they find out about the viewing user's actions and have an adverse reaction to such news. However, when the relationship between content and harm is considered, these individuals are secondary users and individuals as they are not being harmed as a direct consequence of the content existing online; rather, they are impacted by the actions the primary individual has taken due to the content.

3.2 Harms to Third Parties

Third parties become the primary harmed party when there is no harm to the viewing user, and they are instead the first person or persons to be harmed as a consequence of content-promoting behaviours or online instructions for certain actions. Acknowledging that third parties could be harmed in the same manner as viewing users could, in theory, force the consideration of them as a group when platforms assess content that poses a risk of harm.

Third parties are those not overtly considered by the Act and thus not at the forefront of considerations for platforms in adhering to their obligations. These are parties who do not directly engage with hosted user-generated content yet have the potential to be impacted by it via the actions of a viewing user. This can occur when viewing users are influenced by content to engage in specific activities or engage in behaviours not typical of their everyday lives that cause offline disruption, harm or disturbance to third parties. Whilst potentially considered under section 234(5) of the Act,[35] platforms are not obligated to consider these circumstances and impacts, which thus have the potential to continue despite the introduction of safety regulations.
As demonstrated above, there have been multiple recent examples of circumstances where third-party harms can be connected to the existence and proliferation of online content. The accounts of these events have informed the focus of this paper and contributed to the categorisation of harm presented below. Some examples, such as recorded disturbances in schools and educational institutions in 2023,[36] and mass gatherings in St Michael's on Wyre in January 2023,[37] were covered in a BBC exposé

[35] ibid s 234(5).
[36] Sanders (n 22); Wightwick (n 22); Shoesmith (n 22); Busby (n 22).
[37] Andy Gregory, 'Caravan Park Staff 'Told to Lock Doors' as Nicola Bulley Vigilantes Harass Villagers' *The Independent* (15 February 2023) <https://www.independent.co.uk/news/uk/home-news/nicola-bulley-caravan-park-vigilantes-b2281894.html> accessed 10 December 2023; 'Man Held over Nicola Bulley TikTok Post Further Arrested' BBC News (23 June 2023) <https://www.bbc.com/news/uk-england-lancashire-65995807> accessed 11 December 2023; Robyn Vinter, 'Nicola Bulley: Police Issue Dispersal Notices after Social Media Speculation' *The Guardian* (9 February 2023) <https://www.theguardian.com/uk-news/2023/feb/09/nicola-bulley-police-issue-dispersal-notices-after-social-media-speculation> accessed 10 December 2023; College of Policing (n 26).

of the connection between content on TikTok and the actions of viewing users, which adversely impacted and caused harm to third parties.[38]

It is noted at this point that the examples used to produce these three types of harms all resulted in some offline measure being implemented to combat the gathering of individuals and thus reduce the risk of harm, and such intervention only transpired once any harm had already occurred.[39] However, lower-level impacts, such as those discussed below, which affect both third-party individuals and property as part of broader, large-scale disturbances remain unregulated, both online and offline. This suggests that the most effective route to reducing third-party harms stemming from online content would be to recognise them as a variation of 'online harms' arising from user-generated content online, and incorporating this into any regulatory provisions operating within the environment. This would ensure the effective, widespread removal of content containing calls to action and messages of encouragement that have preceded the harm in each set of circumstances.

3.2.1 Economic Harm

The first type of harm experienced by third parties is economic harm. Akin to economic loss established in tort law,[40] these are losses that cannot be physically seen, but can be recognised by a reduction in trade, or actions that cause a third party to lose money.

In the recent examples above such harm has predominantly stemmed from disruption to areas surrounding businesses, which has impacted trade and day-to-day activities. In August 2023, trending content that called for rioting and looting of popular shops on London's Oxford Street saw masses of individuals descend on the area. While, on this occasion, preventative measures were put in place to minimise impact, the police response and the atypical presence of the individuals concerned disrupted everyday activities; for example, shops shut early, causing economic harm.[41] A similar situation arose in late 2024 when farmers drove their tractors into central London in a demonstration against potential tax changes. Popular social media figures spoke out about proposed changes and encouraged people to attend; given the size of the demonstration and the number of vehicles involved, economic harm to surrounding third-party businesses was likely.

Economic harm can also more widespread than to specific shops or third parties, and can be accompanied by other types of harm. Following speculative content on TikTok in January 2023 after a local woman went missing in St Michael's on Wyre, local businesses were subject to speculation and aspersions were cast about them being

[38] Marianna Spring, 'Inside Tiktok's Real-Life Frenzies − from Riots to False Murder Accusations' BBC News (20 September 2023) <https://www.bbc.com/news/technology-66719572> accessed 27 September 2023.
[39] Vinter (n 37).
[40] *Hedley Bryne v Heller & Partners* [1964] AC 465.
[41] Gecsoyler (n 2).

involved in the case. In this instance, not only did individuals face the harms of trespass and police investigations being disrupted, but one local accommodation business was subject to intense speculative content and individuals acting on this content. Whilst specific steps were taken to protect the well-being and safety of residents after a local caravan park was targeted in direct response to online claims,[42] in this instance, such speculation likely caused innocent third parties economic harm due to their businesses being connected to the case in a non-favourable manner.[43]

3.2.2 Physical Harm

There is also evidence that third parties have experienced physical harms due to content being present online. In comparison to both economic and social harms, it is easier in such cases to pinpoint where a third-party is involved. Such harms have been recorded as being a direct result of online content with individuals; for example, when police officers were harmed during demonstrations on Oxford Street,[44] teachers were harmed where educational institutions were subject to student uprising,[45] and multiple individuals were injured, and property damaged, in riots across the UK.[46] In all these situations, the physical harm to individuals was predominantly injuries sustained in resistance to demonstrations brought about by online content rather than viewing users acting in specific ways to injure third parties − the latter of which would now be regulated for under the Act's inclusion of some harms to third parties in its expressed understanding.[47]

Physical harm to third parties has prompted a significant offline reaction, with interventions such as localised responses[48] and dispersal orders[49] being deployed to minimise the future impact on third parties. Notably, in each of our examples of such physical harm, platforms did not remove the encouraging content, in some cases issuing statements of rebuttal that it did not breach service community guidelines.[50] Although many of these instances precede the introduction of the Act and phased implementation of measures, such rebuttals are indicative of a culture across platforms that prioritises removing content where the viewing user might be harmed. This indicates that should content not pose a risk of harm, physical or otherwise, to a viewing user, it will not be removed as part of a platform's regulatory effort. This may change in time as the Act comes into force and invokes duties on platforms to have

---

[42] Gregory (n 37).
[43] Kate Plummer, 'Nicola Bulley: Police Spotted at Caravan Site near Where Dog Walker Disappeared' *The Independent* (13 February 2023) <https://www.independent.co.uk/news/uk/home-news/nicola-bulley-wyreside-farm-caravan-park-b2281352.html> accessed 3 February 2025.
[44] Ashton and Corbishley (n 23); Evans (n 23).
[45] 'Riot, n.' <https://oed.com/dictionary/riot_n>.
[46] Al-Othman et al (n 24).
[47] Online Safety Act, s 234.
[48] Wightwick (n 22).
[49] Shoesmith (n 22).
[50] Busby (n 22).

regard to some third-party impacts. However, the consistent prioritisation of harms to viewing users in both policy direction and narratives means that perhaps the most serious of harms to third parties − those physical in nature − will still occur.

3.2.3 Social Harm

The last type of harm that third parties can experience is social harm. This is more nuanced than physical and economic harms, and can most commonly be seen where the actions of third parties have disturbed and impacted a large number of individuals or communities.

The most prominent example of social harm due to the presence of encouraging or instructive content online is the mass disturbances seen across the UK following the fatal stabbing of three young children in Southport in August 2024. In the immediate aftermath, a post containing misinformation sparked offline unrest, riots and protests as users called upon others to gather in opposition to a falsely informed belief that the suspect was a migrant.[51] This was the highest level of social unrest since 2011, resulting in large numbers of arrests and criminal charges being pressed. The violence and disorder was spread across multiple locations around the UK, and was organised via social media sites and encrypted platforms.[52] In this instance, the predominant harm to third parties was social in nature. Whilst there were reports of isolated physical harm to third parties such as police officers (much like there was in Oxford Street), the spread of pockets of violence across the UK spurred on by online content had a large impact on society overall. Whilst there were arrests and charges on the basis of new offences introduced in the Act,[53] the existence of content promoting or encouraging violence remained present on platforms without moderation. TikTok was used by the police to identify the perpetrators picked up on livestreams[54] On the one hand, online content was actively used to incite and promote violent actions, leading to third-party harms, but on the other, the fact that livestreams could be used to identify criminal activity demonstrates that TikTok as a platform made little effort overall to remove content that had the potential to harm third parties.

Social harm was also seen on a more isolated scale in St Michael's on Wyre in 2023. In this instance, TikTok was retrospectively isolated as the specific platform on which viewing users saw content that was critical in the mass congregation of individuals and harms. In its report, the College of Policing noted the unprecedented social media

[51] Martin, Bancroft and Maddox (n 24).
[52] Spring (n 24).
[53] Holly Evans, 'Woman Arrested over Inaccurate Social Media Post on Identity of Southport Stabbing Suspect' *The Independent* (8 August 2024) <https://www.independent.co.uk/news/uk/crime/southport-stabbings-identity-woman-arrested-police-b2593467.html> accessed 9 September 2024.
[54] Jim Waterson and Vikram Dodd, 'UK Police Monitoring TikTok for Evidence of Criminality at Far-Right Riots' *The Guardian* <https://www.theguardian.com/politics/article/2024/aug/07/uk-police-monitoring-tiktok-for-evidence-of-criminality-at-far-right-riots> accessed 9 September 2024.

attention in the case, highlighting how the specific use of TikTok to spread content about the woman's disappearance resulted in related videos on TikTok gaining 270 million views in total.[55] The unprecedented attention given to the case highlighted the potential for platforms to promote content relating to contemporary unfolding events over other topics. The College commented that users engaged in posting 'speculative content, which may be inaccurate, in the hope of gaining a wider audience'.[56] It also indicated that such content may occur in future due to the behind-the-scenes promotion done by online services.[57] The intense popularity of content that encouraged people undertake their own investigations in St Michael's on Wyre was a main factor in the consequential harm to third parties, reaffirming the role of platforms as a controlling entity where social harm might arise.

4. Discussion – Striking a Balance between Viewing Users and Third Parties

At the core of this paper exists a discussion as to how 'online harms' are defined, identified and mitigated. Despite being the flagship terminology motivating and guiding regulatory efforts over the last five years, there is no universal consensus of what constitutes 'online harm'. This has resulted in isolated definitions of the term being used across jurisdictions, each with its own characteristics and points of reference.[58] Given the lack of consensus on the matter, it is imperative to conduct a contemporary assessment to determine the extent to which third-party harms could be recognised as a potential outcome of problematic content existing online.

Before any potential inclusion to the framework can be considered it is essential to understand to whom harm has to occur for the associated content to be removed as part of the new Act's obligations. Whilst offline responses curtailed some of the impacts within the examples explored above, the most effective way of preventing future third-party harms would be to obligate platforms to recognise third parties as akin to viewing users when making regulatory decisions. In addition to recognising these instances as a variation of 'online harms', the regulatory provisions must also recognise and outline appropriate responses to the role played by the recommendation systems and underlying architecture present on these platforms in influencing the behaviour of users resulting in third-party harms.

4.1 Significance of Regulating Underlying Architecture and Design

Regulatory initiatives place external controls on platforms in connection to the content that they remove or retain by providing descriptors of that which should be

[55] College of Policing (n 26) 5.
[56] ibid 78.
[57] *The TikTok Effect* (n 27).
[58] Benjamin Farrand, 'How Do We Understand Online Harms? The Impact of Conceptual Divides on Regulatory Divergence between the Online Safety Act and Digital Services Act' (2024) *Journal of Media Law* 1.

considered harmful to users and processes that should be followed to reduce the likelihood of users coming to harm. Internally, the underlying architecture and design of platforms can control the content that users see and thus influence the likelihood of economic, physical or social harms occurring. These underlying features can potentially be altered so a user views different topics and levels of potential harm. The possibility of such alterations was raised in connection to factors which drove the prevalence of activities leading to third-party harms in recent years, with former TikTok employees claiming that behind-the-scenes decisions led to such content being pushed to users to increase engagement and boost profits. Whilst this was not the first claim of profit being prioritised over user safety on social media platforms,[59] it does demonstrate the relationship between platform design and choices and the harm evidenced to third parties.

Harm can be exacerbated by platforms shaping interaction and content delivery, as suggested by Price (2021).[60] Conceptualising platforms as social spaces within his wider critique of the draft iterations of the regulations, Price illustrates how it is not only the deficits in the direct provisions concerning harm to users that could mean third-party harms are missed by the regulations, but the understanding and presentation of social media platforms as user-to-user services at the centre of the provisions. Overall, he suggests that the Draft Bill's understanding of platforms as mere hosts of content – now transposed into section 3 of the Act[61] – fails to recognise the social spaces that are created online by platforms in their presentation and curation of content. He argues that by focusing on how individual pieces of content can cause harm, the now-enacted provisions fail to acknowledge the wider adverse impacts that can be caused by how content is delivered to and interacted with by users. Such delivery and interaction are significant when the actuality of third-party harms is considered. The purposeful pushing of content promoting offline action by platforms was indicated to be a root cause of economic, physical and social third-party harms on more than one occasion,[62] the hallmark recommendation systems present being used as a delivery mechanism.[63]

Price's concept of platforms creating social spaces rather than merely being hosts of user-generated content indicates that platforms should be responsible for the space they govern as it is here that user interactions become a product of their environment

[59] Georgia Wells, Jeff Horwitz and Deepa Seetharaman , 'Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show' *Wall Street Journal* (14 September 2021) <https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739> accessed 16 June 2022.
[60] Luke Price, 'Platform Responsibility for Online Harms: Towards a Duty of Care for Online Hazards' (2021) 13 *Journal of Media Law* 238.
[61] Online Safety Act, s 3.
[62] *The TikTok Effect* (n 27).
[63] 'How TikTok Recommends Videos #ForYou' TikTok (16 August 2019) <https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you> accessed 18 December 2023.

– controlled by the platforms themselves – and become either harmful or harmless.[64] The spaces are curated, controlled and governed by the underlying algorithms and architecture, mandating how users congregate in spaces online and the 'reach and outcomes of that congregation'.[65] It can be suggested that responsibility is appropriately assigned as part of regulation, such that consequences of interactions are 'more than the sum of their parts',[66] and introducing the idea that third-party harms can be direct consequences of the congregation of like-minded users online as a result of content prioritisation by platforms.

This further indicates that not only is there a difference between the risk of economic, physical or social harm and their manifestations, but that it is equally important to understand and assess for harm the consequences of content and not just the content itself. Price suggests that by only considering platforms as 'user-to-user services' as per section 3 of the Act,[67] the regulations fail to cover the consequences and social elements of interactions which have been evidenced on multiple occasions to cause harm to users and third parties. Whilst there have been improvements made to the regulations since the first Draft Bill – such as an amendment to the definition of harm to include the dissemination of content, and the mandate that platforms must take into account how 'easily, quickly or widely'[68] an underlying algorithm could spread a piece of illegal content − the focus within the regulations is still on how individual pieces of content present a risk of harm rather than on how social media platforms as a whole can cause harm to individuals. In the opinion of Price (2021), this is a weakness of the whole regime.

Despite their varying locations and timings, the examples of economic, physical and social harms above share the characteristic of being inspired and driven by content created and disseminated on social media platforms. Many of these occurred due to content being shared on TikTok,[69] to which this paper now turns to demonstrate the significance of underlying architecture and design where third-party harms occur within the context of the specific platform.

Whilst TikTok has been noted for its distinct use of personalised recommendation systems,[70] such underlying architecture and design features, which deliver unique content suggestions to users, have now become universal across social media platforms. This personalised content delivery method is underpinned by users engaging with the content they see and spending periods consuming and interacting with the content delivered to them. TikTok's potential to offer infinite content at a

[64] Price (n 60) 239.
[65] ibid.
[66] ibid 255.
[67] Online Safety Act, s 3.
[68] ibid s 9(5)(b)(ii).
[69] Spring (n 38).
[70] Pengda Wang, 'Recommendation Algorithm in TikTok: Strengths, Dilemmas, and Possible Directions' (2022) 10 *International Journal of Social Science Studies* 6066.

fast pace often leads to users finding themselves addicted to the platform and reporting adverse impacts in their lives.[71] Whilst each of our examples centred on TikTok offering recommendations to users, which drove their offline actions,[72] such features are not unique to the platform. Harm to viewing users and third parties is not an isolated issue, but one that must be addressed across the ecosystem of social media platforms more widely.

Wang et al (2022)[73] previously investigated the manner in which users respond to content delivered to them via recommendation algorithms on TikTok in particular. When consulting young people about any healthy lifestyle choices they made following such interactions, the authors identified four broad categories of consequential actions: immediate actions; planning actions; planned actions; and reflective actions. This categorisation is relevant when determining where harm to third parties can occur. When considered in line with the potential for users to engage in behaviours with the potential to adversely impact others, planned actions are perhaps the most significant and can be identified within the above examples.

Termed by the authors as 'actions [users] take as a consequence of watching food content on TikTok',[74] planned actions are those that occur immediately after exposure, or engagement to or with, a piece of content on the platform. For the purposes of this paper, this means that responding with comments or 'likes' to a piece of content promoting offline actions, such as sit-ins or protests, could be recognised as immediate actions. Whilst not physical manifestations of harm, these immediate actions promote further engagement and the popularisation of problematic content with the potential to cause harm, and encourage others to partake in the actions that have been shown to cause harm to third parties. Planning actions are termed by Wang et al as encapsulating the 'ongoing decision-making process [their] participants reported in response to TikTok's food content'.[75] This represents the user's thought processes, feelings and reactions after seeing or interacting with a piece of content but no longer engaging with the content itself. For the purpose of this paper, planning actions are plans made by viewing users to attend or engage in actions such as protests or sit-ins. Whilst not harmful to them as primary viewers, these actions could then lead to third-party harms.

[71] Troy Smith and Andy Short, 'Needs Affordance as a Key Factor in Likelihood of Problematic Social Media Use: Validation, Latent Profile Analysis and Comparison of TikTok and Facebook Problematic Use Measures' (2022) 129 *Addictive Behaviors* 107259.
[72] Spring (n 38).
[73] Chun-Han Wang et al, '"TikTok Made Me Do It": Teenagers' Perception and Use of Food Content on TikTok' IDC '22: Proceedings of the 21st Annual ACM Interaction Design and Children Conference (27 June 2022) 458 <https://dl.acm.org/doi/10.1145/3501712.3535290> accessed 28 November 2023.
[74] ibid 461.
[75] ibid 462.

So, planned actions refer to the carrying out of any ideas that viewing users have had in response to the content they have engaged with and consumed; this is where the potential for harm to third parties materialises and can be evidenced. Participation in sit-ins and looting, damage to property, injuries to individuals, and trespassing on private property can all be recognised as planned actions causing harm to third parties. In their research into healthy lifestyle choices Wang et al explain that these actions are undertaken as a consequence of users 'remembering specific TikTok content, whether implicitly through recollection or explicitly through scrolling through their saved recipes, and putting these corresponding plans to action'.[76] This shows a direct correlation between the content that is hosted online and the actions users take that, in this instance, cause harm to third parties that otherwise would not have occurred. It also affirms claims made concerning the influence of TikTok in connection to demonstrations, and supports the idea that the underlying architecture on the 'For You' page has a role to play in the delivery of content that would ultimately lead to harm to third parties and property.[77] It thus affirms the previous suggestions that both the role of platforms and the presence of potentially problematic content need to be acknowledged to limit the potential for third-party harms to occur.

4.2 Deficits in Regulation

When the UK's new online safety regulations were drafted and adopted, the focus was primarily with the individual viewing users. They did not cater for third parties being harmed due to the existence of this online content – a reality that is increasingly common.[78] Whilst the importance of keeping viewing users safe in the first instance is not undervalued within this work, it does indicate a deficit in the regulations more widely.

4.2.1 Understanding 'Harm' in the Regulations

Harm as a concept has developed within the confines of the provisions, being both added to and subtracted from as the regulations have developed. Lobbying by various groups of charities has lead to the successful inclusion of provisions reflecting contemporary issues not foreseen in the first White Paper,[79] such as the criminal liability introduced in section 181 for persons engaging in threatening communications with the intent to harm others.[80] These developments indicate that, going forward, there is potential for harm to third parties to be recognised as a legitimate by-product of online content and be treated as akin to harm to viewing users.

As previously explored within this paper, section 234 defines outcomes that should be considered 'harmful', the impacts of which are seen as either 'physical or

[76] ibid.
[77] *The TikTok Effect* (n 27); Spring (n 38).
[78] Department for Science, Innovation and Technology et al (n 7).
[79] The Department for Digital, Culture, Media and Sport (n 9).
[80] Online Safety Act, s 181.

psychological'.[81] This definition is arguably the regulatory representation of the 'online harms' phenomena. During the drafting stages, this was previously defined by stakeholders as 'user generated content or behaviour that is illegal or could cause significant physical or psychological harm to a person'.[82] Indicating that both behaviour and content can create harm, but with no reference to by whom harm has to be experienced beyond 'a person',[83] the possibility that either viewing users or third parties could be harmed was introduced. However, as this definition has only been provided in relation to the Draft versions of the regulations, the requirement for platforms to consider the potential impacts of content on third parties is reduced in the currently enacted provisions.

4.2.2 Section 234(5)

For a framework deeply rooted in the idea of reducing harm,[84] the definition provided is brief compared to the various reported consequences users have previously stated they feel are harmful. Section 234(5) acknowledges that harm can occur to individuals other than the viewing user, stating that references to harm within the regulatory framework should include considerations of circumstances where 'as a result'[85] of seeing or interacting with content, an individual does or says 'something to another individual that results in harm',[86] or 'increases the likelihood of such harm'.[87] Whilst this provides some protection and consideration of potential third-party harms, the subsection lacks specific limits that could pose significant challenges for platforms that seek to adhere to the new obligations. Moreover, this provision only applies to individuals, meaning that businesses, such as those harmed in the Oxford Street disturbances, would not be included in the considerations of platforms when determining if content poses potential harm.

Firstly, the provision does not indicate a maximum number of people that could be impacted by a user acting on content. Due to this, potential third-party harms stemming from user-generated content between one additional individual to endless additional individuals could weaken the intended impact of the inclusion of such a provision. It is possible that this provision has been included in the Act to allow for instances such as content encouraging a user to bully someone else offline to be accounted for and reduced, as such an example would see no harm to the viewing user in the first instance, but could see harm to a third party if the viewing user acted on the content presented to them. This could, theoretically, address where significant

---

[81] ibid 234(2).
[82] 'Understanding and Reporting Online Harms on Your Online Platform' GOV.UK <https://www.gov.uk/guidance/understanding-and-reporting-online-harms-on-your-online-platform> accessed 18 December 2023.
[83] ibid.
[84] The Department for Digital, Culture, Media and Sport, *Online Harms White Paper: Full Government Response to the Consultation* (2020) para 2.11.
[85] Online Safety Act, s 234(5)(b).
[86] ibid.
[87] ibid.

events leading to third-party harms have occurred. However, with no limitation on the number of individuals, platforms may be unable to characterise content as harmful or non-harmful in this respect and, therefore, the content would not be removed.

Secondly, section 234(5) appears to exclude the need for platforms to consider any harm to third-party property whether content is deemed harmful or non-harmful.[88] It is acknowledged that the Act was introduced to tackle adverse outcomes of interactions that individuals have with content, and thus property damage or economic harm may not have been considered by policymakers. However, this presents a gap in the legislation to reduce situations where the presence of online content has led to damage to third-party property through the actions of viewing users.

4.2.3 Section 62(4) – Primary Priority Content Harmful to Children

Section 62(4) of the Act states that a piece of content that encourages, promotes or provides 'instructions for an act of serious violence against another person'[89] is 'priority content that could be harmful to children'. Platforms must therefore remove the content to satisfy their obligations to protect children online. 'Serious violence'[90] arguably represents an extreme manifestation of actions arising from the viewing of a piece of content. This means that the threshold of harm that content would have to reach is high, presenting the possibility that minor disturbances and instances of violence occurring as a consequence of viewing and acting upon content would not have to be determined as potentially harmful to children as required under section 60.[91] Yet, the expansion does appear to recognise the possibility of harm to other individuals beyond the viewing user as a consequence of content, which would allow harm to third parties to be recognised and conceptualised within the context of the regulations. This would quell future frenzies as the content would be removed in the first instance because it poses a potential risk of harm to third parties.

However, within the examples of third-party harms, both calls to congregate and offline action beyond calls for violence have been evidenced, meaning that content that results in impacts such as reputational damage or offline interference with police investigations would not be covered. At this point, an interrogation of what is meant by 'violence' in this regulatory context is worthwhile. In the absence of a legislation-specific definition,[92] it can be suggested that 'serious violence' is likely to cover public order offences, like riots, violent disorder and affray, but unlikely to cover fear of harassment. This means that examples such as the demonstrations witnessed in schools against toilet policies,[93] are unlikely to meet the threshold for 'serious

[88] ibid s 234(5).
[89] ibid s 62(4).
[90] ibid.
[91] ibid s 60.
[92] ibid s 236.
[93] See Busby (n 23).

violence'.[94] Furthermore, section 62(8) of the Act expands the definition of priority content harmful to children as 'content which encourages, promotes or provides instructions for a challenge or stunt highly likely to result in serious injury to the person who does it or to someone else.'[95] This suggests that should there be a possibility that an individual beyond the viewing user suffers 'serious injury'[96] as a consequence of interaction with a piece of content, a platform should automatically determine this as being harmful to children.

Whilst section 62 is a useful interpretation of the provisions given the potential of harm to third parties being accounted for within the regulations. Whilst services are obligated to have regard for such impacts, these only apply to content that is likely to harm children, indicating two things. Firstly, a provision that relates to content that is likely to harm children may lead to circumstances where platforms and services only consider third-party harms that occur to children, rather than those that occur to those aged over 18 years of age. Secondly, if section 62 is advanced as a standard for assessing harmful content that poses a potential risk to third-party children, this would establish a higher standard than if there were a separate provision focusing solely on potential harms to adults.

However, for there to exist a successful framework that reduces harm arising from online content in its various forms − both for viewing users and third parties − the most effective way forward would be to introduce new specific guidance or regulatory provisions that impose additional obligations on platforms. This would ensure that interpretations or classifications of specific user groups do not diminish the potential for all harm types and examples to be mitigated in line with the Act's regulatory aims.[97]

5. Recommendations and Ways Forward

At the time of writing, the farmer protests witnessed across the UK − an example of content being linked to offline actions impacting third parties – are ongoing. The Act and subsidiary Ofcom Codes are also being phased into force, signalling a new era for online safety.[98] That these two realities exist in parallel demonstrates the ongoing risk of harm posed by the ability of online content to change the behaviour of individuals. Therefore, there is an increasing need to recognise and establish the regulatory path forward to reduce situations where third parties are placed in a position of potential harm due to content hosted online. This paper presents two pragmatic ways forward,

[94] ibid s 62(4).
[95] ibid s 62(8).
[96] ibid.
[97] Department for Science, Innovation and Technology et al (n 6).
[98] Ofcom, 'Implementing the Online Safety Act: Progress Update' 14 <https://www.ofcom.org.uk/siteassets/resources/documents/online-safety/information-for-industry/roadmap/2024/ofcoms-approach-to-implementing-the-online-safety-act-2024.pdf?v=383285> accessed 17 October 2024.

firstly introducing the potential for regulatory acknowledgement of third parties, and then revisiting the importance of regulating underlying design choices and architecture as influential factors in third-party harms.

5.1 Clarification of the Limits of Harm and Acknowledgement of Third Parties

Clarifying in further guidance the limits of harm expressed in the regulations is perhaps the most straightforward way to establish the equivalence of third-party harms to viewing user harms within the online safety regulations. The phased introduction of Ofcom's subsidiary safety codes, which is expected to continue throughout this new regulatory era, presents an opportunity to make necessary clarifications and expansions.[99]

Throughout, this paper has suggested that because the Act focuses on the individual viewing user, it fails to recognise that harm stemming from online content and interactions is nuanced and can occur to third parties beyond the scope of the provisions. Therefore, going forward, it is key that Ofcom projects and maintains an all-encompassing understanding of harm beyond that set out in section 234 to ensure that harm in its many forms is reduced, and that the UK meets its goal of being the safest place to be online in the world.[100]

To do this, the definition of 'physical or psychological'[101] present in section 234 should be reviewed, and expanded on within the planned secondary guidance or the explanatory notes to provide more depth. The current definition may have been left vague and non-prescriptive beyond general descriptors to allow future-proofing. However, for the purposes of pragmatic application, as online platforms and services develop and evolve, a wider definition is more beneficial to platforms in their day-to-day monitoring and moderation of content.

5.2 Utilising Existing Legal Mechanisms to Recognise Third Parties

To inform any expansions to regulatory guidance, the tortious origins[102] of the safety regulations could provide policymakers with a pragmatic way to allow for the acknowledgement of harm to third parties without considerations being overwhelming for platforms. The *Alcock*[103] control mechanisms of primary and secondary victims provided by the courts in connection with traumatising events that occurred offline but were broadcast on television, extending the potential list of entities impacted, could be echoed here.

---

[99] ibid.
[100] Department for Science, Innovation and Technology et al (n 6).
[101] Online Safety Act, s 234(2).
[102] The Department for Digital, Culture, Media and Sport (n 7); Lorna Woods and William Perrin, 'Online Harm Reduction – a Statutory Duty of Care and Regulator' (Carnegie Trust UKE, 2019) <https://d1ssu070pg2v9i.cloudfront.net/pex/carnegie_uk_trust/2019/04/08091652/Online-harm-reduction-a-statutory-duty-of-care-and-regulator.pdf>.
[103] *Alcock v Chief Constable of South Yorkshire Police* [1992] 1 AC 310; *Paul, Polmear and Purchase and another v Royal Wolverhampton NHS Trust* [2024] UKSC 1.

Following the 1989 Hillsborough disaster, numerous claims were made for damages to compensate for the harm individuals experienced from seeing the events unfold. The courts decided to impose a limitation on who could be recognised as a victim of the event. Primary victims were those present on the day of the disaster, while secondary victims included friends and family who suffered from witnessing the events through live media coverage. Any individual seeking compensation needed to fit into one of these categories, restricting the number of parties that might have been harmed. Given the relevance of the *Alcock* doctrine in new media,[104] similar controls could be put in place in regard to the harmful impacts of online content, in order to acknowledge such impacts whilst also providing practical limitations that prevent platforms from being overburdened with considerations when assessing content. This would recognise the freedom of speech and freedom to organise and assemble protections afforded to individuals, such as those congregating against tax changes for farmers, whilst balancing obligations owed by platforms to protect people from harm.

In the context of online content and following *Alcock*, individual viewing users could be considered akin to primary victims (ie, as those potentially harmed in the immediate aftermath of events occurring or being witnessed), with third parties being akin to secondary victims. For a platform or service to consider a third party and decide whether content remained or was removed on these grounds, there would have to be a significant connection to the topic, instructions or likely outcomes of the user-generated content being assessed.

To see how this might work in practice, we can apply it to the example of St Michael's on Wyre. Under the proposed categorisations, primary parties would be those who interacted with content in the first instance, and who were not overtly harmed by seeing it, but were subsequently arrested or banned from the area.[105] Meanwhile, the individuals who lived in the village at the time who faced adverse attention from users online would be categorised as secondary parties. In accordance with the contemporary expansions of the *Alcock* distinctions, these parties would need to demonstrate that they were in the same time and space of an event caused by the viewing user acting on user-generated content.[106] This means that those who were victims of trespass or injury due to the presence of viewing users could qualify, but others perhaps living close by but not directly impacted would not. This would impose a workable limit and provide guidance on the considerations platforms have to make in relation to third parties, refining the focus to those who are likely to be impacted, making the acknowledgement of third parties potentially workable in the regulatory environment.

---

[104] Bela Bonita Chatterjee, 'Rethinking Alcock in the New Media Age' (2016) 7 *Journal of European Tort Law* 272.
[105] 'Man Held over Nicola Bulley TikTok Post Further Arrested' (n 37).
[106] *Paul, Polmear and Purchase and another v Royal Wolverhampton NHS Trust* (n 103).

5.3 Recognition of Recommender Systems as a Basis for Regulation

The second way in which the harms to third parties and property could be reduced is to regulate the internal mechanics of the platforms on which the problematic content is shared and seen by users. Across the examples featured in this work and the types of the harms that have stemmed from them, there were indications of underlying platform features, such as personalised feeds and recommender systems, purposefully promoting content to users who were likely to take action after viewing it, indicating that, behind the scenes, platforms were making decisions that increased the likelihood of harm occurring.

This paper builds on previous critiques concerning the role of underlying architecture in building harmful environments online and suggests that in order to curtail third-party harm, platforms should be obligated to be transparent about their decision-making (automated and otherwise) on how content is pushed and promoted to users. Gillespie (2022) suggested such a remedy. He explored how using reduction mechanisms can co-exist with, and provide an alternative to, the removal of content online.[107] In this system, content stays on the platform and can be found by users who search for it, but it is not actively recommended. This approach helps platforms regulate content that does not directly violate community guidelines but could still cause harm or lead to negative discussions among users and any consequential third-party harms. Should such reduction measures have been utilised or mandated in connection to the third-party harms showcased in this work, the harm might have been significantly reduced, as these were arguably reliant on the spread of users engaging and acting on the instructions provided.

Mandating such measures could significantly reduce the prevalence of offline third-party harms to people and property overall. This paper has demonstrated that where third-party harms occur, the catalyst user-generated content has been viral in nature, being seen by numerous viewers and prompting users to take offline action.[108] Providing an alternative method of moderation to services (as opposed to the on-/off-platform binary currently expressed in regulations) would force both the acknowledgement of underlying recommender systems and provide an avenue of moderation that enables users over the age of 18 (the Act has different provisions for the protection of children) to interact with content where it is safe to do so. Of course, there is the potential for users to actively seek out content that encourages actions with the potential to harm others, as was seen with both the evidenced educational disruptions,[109] and those attending St Michael's on Wyre. However, this is where the broadening of what constitutes 'harm' for the purposes of regulation and the acknowledgement of third parties as individuals who can be harmed is welcomed to work in tandem with other regulatory measures.

---

[107] Tarleton Gillespie, 'Do Not Recommend? Reduction as a Form of Content Moderation' (2022) 8 *Social Media + Society* 1.
[108] Spring (n 38).
[109] Sanders (n 22); Wightwick (n 22).

6. Conclusions and Further Works

This paper has shown that the current online safety regulations protecting people who are harmed as a consequence of user-generated content existing online are predominantly aimed at the harms caused to individual viewing users interacting with content at a primary level, and leaves other manifestations of harm largely unrecognised and unregulated. By exploring the types of harm third parties experience in parallel to viewing users, it has been shown that user-generated online content can also manifest harm in ways that are not accounted for in regulations. This has called into question the conceptualisation of the phenomena and characteristics of 'harm' within this regulatory context and opened up discussions as to how regulatory frameworks can best adapt and align with the reported offline circumstances.

It has been suggested throughout that in order to successfully meet the promise made of the UK's online safety regulations – that the UK will become the safest place for users to be online in the world[110] – the legislation needs to be either updated or supplemented with by subsidiary codes. An expanded definition in section 234 of the Act of that which constitutes harm and to whom harm can be caused would provide for platforms considering impacts to third parties when discharging their new duties and take steps towards the framework achieving its stated goal. Recognition of how underlying design choices and how architecture platforms operate and present content influences and normalises the actions of viewing users that then go on to harm third parties could lead to fewer instances of harm stemming from online content overall.

This paper is not exhaustive in its coverage or conclusions, but presents avenues for further inquiry. There are two main points to note. Firstly, as this paper has relied on second-hand accounts and reports of disturbances to interpret and recognise third-party harms to property and people, there is scope for further research to definitively establish the causal relationship between user-generated content hosted online and third-party harms. Secondly, it has initiated discussions of what it means to tackle online harms in the UK in the new regulatory era brought in by the Act. It intentionally leaves unanswered the question of the contemporary meaning of 'harm' and 'online harms'. As regulations are implemented, further research is needed to define more precisely these concepts, especially as platforms strive to fulfil their responsibilities and user interactions with content change.

[110] Department for Science, Innovation and Technology et al (n 6).