The Lost Clause - Exploring the potential impact of amendments to the definition of harm to children in the UK's Online Safety Bill

Ellie Colegate
School of Law, Horizon Centre for Doctoral Training, University of Nottingham, Nottingham, United Kingdom
ellie.colegate@nottingham.ac.uk

Abstract

Introduced into Parliament in March 2022, the Online Safety Bill[1] is intended by the government to be a regulatory instrument that will make the UK 'the safest place in the world to be online'[2]. However, as the latest publication in a set of regulatory proposals and policy documents, the specific regulatory tools contained therein - the digital duties of care - have been subject to debate and development. Both the Online Safety Bill and its draft predecessor contain provisions imparting responsibilities and obligations onto services that host user-generated content to have regard for and take action against potentially harmful content. Whilst the provisions are intended to impact all users who interact with platforms that host user-generated content, specific clauses are dedicated to reducing potentially harmful content to children and young people. The impact of harmful content on young people has been well documented across both industry and academic publications, confirming that this is a contemporary issue facing companies and governments alike. Yet, the Draft Bill published in May 2021[3] and the Bill finalised before being passed to the Lords in December 2022 differ in their interpretations and scoping of what content could be considered harmful to children, with the latter adopting a tighter yet unclear scope. This exploratory paper will assess the two sections covering content that could be harmful to children in both the May 2021 Draft Bill and December 2022 Draft Bill, comparing the specific wordings, mapping the developments, and presenting the potential issues that this change in scope could have in practice for the reduction of harmful content online.

Introduction

Since 2019, the UK government has been developing and refining its Online Safety legislation. Starting with the Online Harms White Paper and recently introducing the Online Safety Bill into Parliament for debate, the proposed regulations seek to improve user safety online by bolstering provisions to reduce potentially harmful content online. However, delays and discussions around the contents of the Bill in its entirety have seen substantial changes to definitions and scoping of provisions key to regulation, opening up the debate to assess which version of the Bill, introduced publicly in March 2022 or the updated version from December 2022 should be taken forward.

[1] Online Safety Bill 2022.
[2] The Department for Digital, Culture, Media and Sport, *The Online Harms White Paper* (2019) 4.
[3] Draft Online Safety Bill 2021.

As part of their adherence to overarching 'duties of care,' online platforms, such as popular social media platforms, must check and monitor the content they host to ensure user safety and minimise the risk of harm online.[4] The provisions in both versions of the Bill mandate that services should conduct specific scoping exercises to determine the aspects of their service that children could access.[5] Platforms are expected to carry out and maintain risk assessments concerning users designated as children[6], adhering to and complying with safety duties to mitigate and manage the risk of harm to the user group, operating proportionate systems and processes that should, theoretically, prevent children from encountering content that could be potentially harmful to them.[7] This identification of potentially harmful content is vital to the effective identification, management, and eventual reduction of harmful content overall. Therefore, the provisions also provide descriptors of content that should be considered harmful to children[8] and the definition of harm within the context of content regulation under the online safety regime.[9] The workability, and thus reduction of content potentially harmful to children, are ambitions set by the regulations that have attracted criticism from charities and academics indicating that provisions introduced are potentially too vague for effective regulation[10], yet have the potential to overregulate user-generated content impacting fundamental rights such as privacy and free speech.[11]

However, there have been notable changes to the provisions as a whole, with those first publicly introduced in the May 2021 Draft Bill being developed and adapted to the extent that those now contained within the December 2022 Draft Bill now represent a set of regulations, that in if enacted as currently written could be problematic for the identification and reduction of content harmful to children.[12] It is these differences that this paper will be focusing on, presenting the potential for these to be problematic when juxtaposed with reported experiences of children during the time that the provisions have been developing. Concentrating on the removal and amendments to descriptors that would aid platforms in the identification and the thresholds contained therein, this paper will illustrate that the provisions currently contained within the December 2022 Draft Bill have been developed and adapted to such an extent that their overall workability and intelligibility have been lessened and could have significant consequences for children operating on these platforms and their safety online.

To address these goals, this paper will explore the changes that have occurred between the May 2021 Draft Bill and the December 2022 Draft Bill, focusing specifically on sections defining content that could be harmful to children and the legislative definition of harm. Firstly, there will be a summary of the background to these provisions overall and the changes that have occurred between May 2021 and December 2022 and the criticisms of the Bill thus far. Secondly, the main arguments of this paper will then place the new December 2022 sections, those currently set to be enacted and influence regulation in real time, with reported experiences of this demographic as a user group to show where problems are likely to occur.

[4] The Department for Digital, Culture, Media and Sport (n 2) 7.
[5] Online Safety Bill s 11(13).
[6] ibid 10.
[7] ibid 11.
[8] ibid 53.
[9] ibid 187.
[10] Laura Higson-Bliss, 'Online Safety Bill: Ambiguous Definitions of Harm Could Threaten Freedom of Speech – Instead of Protecting It' *The Conversation* (22 March 2022) <https://theconversation.com/online-safety-bill-ambiguous-definitions-of-harm-could-threaten-freedom-of-speech-instead-of-protecting-it-179514> accessed 1 July 2022.
[11] Jim Killock, 'Internet Policy Is Broken' (*Open Rights Group*, 10 March 2022) <https://www.openrightsgroup.org/blog/internet-policy-is-broken/> accessed 1 July 2022.
[12] Joe Woodhouse, 'Analysis of the Online Safety Bill' (House of Commons Library 2022) 9506.

Thirdly, as a set of regulations goes through the Parliamentary process, the planned progression will be considered to highlight ways forward. After that, a statement for the scope of further works and conclusions will occur.

Background to Regulation

As a self-contained set of provisions, the current complied Draft Bill published in December 2022 has undergone various amendments and changes. Evolving from an abstract idea of regulation[13] to provisions that, if enacted, should be able to regulate platforms in real time. These provisions have been developed with the aim of combating problematic content sharing online, causing harm to individual users, with pressure being placed on the UK Government to take action to protect users online from harm and the risk of harm.

The main objective of the regulation's, stemming from the 2019 White Paper and recognised in the December 2022 Draft Bill under assessment here, is to enhance the safety of internet users by regulating content on platforms termed 'user-to-user services'[14]. These services are those by which content can be shared from one user to another or multiple users, which can be encountered via another user.[15] Whilst there is no definitive list of which commonplace platforms will be considered user-to-user services, there is a consensus that popular social media platforms such as Facebook, Instagram, TikTok and Twitter will be subject to provisions due to them all offering users the ability to create, upload and share content as the predominant method of platform engagement.

This objective has been maintained throughout the development process and can be evidenced in the various iterations of the regulations; however, whilst the framework aims to 'improve user safety'[16], it notably will not 'eliminate harm or risk of harm entirely.'[17] Indicating that provisions will provide parameters for identifying content to reduce the overall harmful content, but their implementation cannot guarantee that services will lack harmful content entirely. This is notable considering the particular attention the regulations have given to children as users and the specific provisions services have to adhere to ensure user group interactions remain harm minimal. Under the December 2022 Draft Bill, platforms considered to be category 1 services – a 'regulated user-to-user service'[18] determined to meet threshold conditions outlined by the regulator – must illustrate that they have carried out a risk assessment of content in relation to child users where the user group is likely to access their services[19], update the risk assessment in relation where significant changes are made to the risk profile by the regulator[20], and identify 'non designated

[13] Lorna Woods, 'The Duty of Care in the Online Harms White Paper' (2019) 11 Journal of Media Law 6; 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>.
[14] Online Safety Bill (As Amended on Report) 2022 s 2.
[15] ibid.
[16] The Department for Digital, Culture, Media and Sport, *Online Harms White Paper: Full Government Response to the Consultation* (2020) para 2.11.
[17] ibid.
[18] Online Safety Bill (As Amended on Report) s 85(3)(a).
[19] ibid 10(2).
[20] ibid 10(3).

content that is harmful to children'[21], and take action to 'mitigate and manage'[22] the risk of harm to children to adhere to safety duties to illustrate that they adhere to their overall duties.

Under both the risk assessment and safety duties, it is vital that there is the successful identification of content harmful to children. This is critical to the effective risk regulation, management and eventual reduction of harmful content. Therefore, the provisions provide descriptors of content that would be considered harmful to children[23] and defines harm within the context of content regulation under the online safety regime.[24] All of which have been developed since the initial Online Harms White Paper summarised plans for the UK to become 'the safest place in the world to be online.'[25]

However, these have been subject to criticism throughout, especially concerning the clarity of provisions and the theoretical protection offered to children. As a user group, evidenced on multiple occasions in policy and research, harmed by the content they consume online[26], such is key to adequately addressed within regulations.

The regulations themselves have been subject to much criticism, with attention given at a White Paper stage to the lack of legal basis on which the regulations would have been based. Woods[27], Phippen and Bond[28], Rowbottom[29], and Nyamutata[30] all address within their contributions, the latter of which is particularly pertinent to this paper as it also focuses on children as a user group. Within their contribution, Nyamutata indicated that when the regulatory proposals are placed in the broader legal context, comparisons could be drawn between the digital duties of care – as the regulations were previously termed – and the traditional duties of care evidenced in Tort Law. They suggest that when the digital well-being of children is considered, comparisons must be made to pre-existing duties of care to assist successful content regulation and maintenance of well-being. This paper partially departs from this understanding as the developments of the regulations have indicated that these are distinct new duties. However, it does recognise merit in

[21] ibid 10(6)(b)(iii).
[22] ibid 11(2)(a).
[23]Online Safety Bill s 53.
[24] ibid 187.
[25] The Department for Digital, Culture, Media and Sport (n 2) 4.
[26] OFCOM, 'Online Nation Report 2021' (OFCOM 2021) <https://www.OFCOM.org.uk/__data/assets/pdf_file/0013/220414/online-nation-2021-report.pdf>; Sonia Livingstone, 'Risk and Harm on the Internet' in Amy B Jordan (ed), *Media and the well-being of children and adolescents* (Oxford University Press 2014); Vera Slavtcheva-Petkova, Victoria Jane Nash and Monica Bulger, 'Evidence on the Extent of Harms Experienced by Children as a Result of Online Risks: Implications for Policy and Research' (2015) 18 Information, Communication & Society 48; Deborah Richards, Patrina HY Caldwell and Henry Go, 'Impact of Social Media on the Health of Children and Young People: Social Media and the Health of Young People' (2015) 51 Journal of Paediatrics and Child Health 1152; Paul Best, Roger Manktelow and Brian Taylor, 'Online Communication, Social Media and Adolescent Wellbeing: A Systematic Narrative Review' (2014) 41 Children and Youth Services Review 27; Calli Tzani and others, 'A Description and Examination of Cyber-bullying Victimisation in the UK' (2021) 18 Journal of Investigative Psychology and Offender Profiling 157; Nina Jacob, Rhiannon Evans and Jonathan Scourfield, 'The Influence of Online Images on Self-Harm: A Qualitative Study of Young People Aged 16–24' (2017) 60 Journal of Adolescence 140.
[27] Woods (n 13).
[28] Andy Phippen and Emma Bond, 'The Online Harms Spearmint Paper - Just More Doing More?' (2019) 30 Entertainment Law Review.
[29] Jacob Rowbottom, 'Introduction' (2019) 11 Journal of Media Law 1.
[30] Conrad Nyamutata, 'Childhood in the Digital Age: A Socio-Cultural and Legal Analysis of the UK's Proposed Virtual Legal Duty of Care' (2019) 27 International Journal of Law and Information Technology 311.

exploring these in line with the traditional knowledge in an attempt to gain clarity where terms are shared between the two areas of law.

There has also been criticism of the provisions and their impact on human rights, such as freedom of expression, with Harbinja and Leiser suggesting that enacting the regulations as they were drafted in March 2022 would bring about circumstances in which content is unjustly removed where platforms regulate with caution rather than accuracy.[31] Whilst beyond this paper's focus, these are key to note as they illustrate the difficulty of regulating such contemporary, new areas such as online platforms in the interests of various stakeholder parties such as governments, platforms, regulators, and user groups, such concerns also highlight the need for clear and precise regulations – a critique this paper share for an alternative rationale.

Most recently, Trengove et al. have highlighted notable issues with the scope of the regulations, indicating that such is too broad with its powers and may burden platform providers/services.[32] A suggestion that this paper will support in relation to platforms offering services to children, further illustrating that the issues highlighted by Trengove et al. contained within the intermediary March 2022 Draft Bill – that published between the May 2021 Draft Bill and the December 2022 Draft Bill – remain in the most recent iteration.

Overview of changes – Criteria for Content Assessment

Despite criticism and debates, developments and changes have occurred between the May 2021 and December 2022 Draft Bills, with the latter presenting specific provisions that retain the potential to be problematic and ineffective in practice. Section 54 – that which outlines the content that should be considered harmful to children – and Section 201 – the legislative and regulatory definition of harm – both have key positions and consequences for regulating children's interactions online. Yet, both have specifics and elements remaining unclear and undefined which could cause issues in practice and hinder the goal of the regulations to reduce harmful content.

There have been three publications of the above provisions, with the most recent in December – that of analytical focus in the discussion – remaining free from substantial changes. In the interests of contemporary accuracy, this paper will compare the December 2022 Draft Bill with the May 2021 Draft Bill to assess the extent to which these provisions have been altered and the consequences of these changes on any eventual regulation. Numerically, there have been changes to the positioning of provisions as new amendments have been proposed and adopted. However, this is not of interest to this paper, and as such, the designations presented in the December 2022 Draft Bill will be those taken forth.

The May 2021 Draft Bill marked the first legislative iteration of the online safety regulations, introducing for the first time the specific sections and wording mandating the duties platforms and companies will have to adhere to. Outlining for the first-time content that would be considered harmful to children, Section 45 stated that content would be considered harmful if a platform or service if such had previously been determined either primary priority or priority content by the Secretary of State or if a platform had 'reasonable grounds to believe that the nature of the content is such that there is a material risk of the content having, or indirectly having, a significant adverse physical or psychological impact on a child of ordinary sensibilities ("C").'[33] Within the December 2022 Draft Bill, the discretion granted to the Secretary of State to determine both primary priority and priority content remains. However, there have been significant alterations to the descriptor given to platforms. The December Bill states content will be harmful if it has

[31] 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.
[32] Markus Trengove and others, 'A Critical Review of the Online Safety Bill' (2022) 3 Patterns 100544.
[33] Draft Online Safety Bill s 45(3).

been determined to be primary priority[34] or priority content[35] by the Secretary of State or if such is not covered in by either of the content descriptors yet poses a 'material risk of significant harm to an appreciable number of children in the United Kingdom.'[36]

As a term key to online safety regulations overall, the definition of harm contained within the provisions has also been developed. The May 2021 Draft Bill previously stated that harm would refer to either 'physical or psychological harm'[37]. The December 2022 Draft Bill has expanded on this and provides a more expansive definition yet maintaining the precursor of such being physical or psychological. Section 201 now considers how a piece of content is disseminated, indicating that harm to users can occur from the nature of the content itself[38], how it is disseminated to users[39], and the manner of its dissemination.[40] The new definition of harm encompasses instances where content is 'repeatedly sent to an individual by one person or different people'[41] indicating that platforms should now aim to look beyond the nature and subject of content and consider the contextual elements present on their platform to determine if such presents harm to children.

These changes have occurred following structured consultations and Select Committee discussions, indicating that charities should welcome these developments and those consulted concerned with protecting children online. However, in line with the provisions as a set, these have attracted criticism, with suggestions being made that these provisions still need to go further to protect children in real time.[42] This paper acknowledges such comments and will now adopt a broader perspective of the May 2021 and December 2022 Draft Bills to assess and suggest any further appropriate changes to increase the likelihood of harm reduction for children by comparing such with the reported experiences of the user group.

The Revised definition of harm – clear on paper, unclear in practice?

The successful identification of the content that harms children is a critical prerequisite to reducing such online and protecting users as a direct consequence. Therefore, it is arguably vital for all stakeholders – users, platforms, and regulators – to adequately and clearly define what is meant by harm within this specific regulatory context.

The December 2022 Bill, as summarised above, defines harm as either 'physical or psychological'[43] yet recognises that how user-generated content is disseminated can also give rise to harm.[44] This development introduces multiple scenarios that platforms must consider before conclusively determining content to be harmful, making such potentially unworkable in practice. When the reported experiences of children and young people are considered, this expansion and the current wording of the provision are welcomed. It is

[34] Online Safety Bill (As Amended on Report) s 54(a).
[35] ibid 54(b).
[36] ibid 54(c).
[37] Draft Online Safety Bill s 137.
[38] Online Safety Bill (As Amended on Report) s 201.
[39] ibid 201(3)(b).
[40] ibid 201(3)(c).
[41] ibid.
[42] The British Psychological Society, 'Online Safety Bill Still Leaves Children and Young People Vulnerable to Harm and Must Be Strengthened, Warn BPS and YoungMinds | BPS' (*The British Psychological Society*, 19 April 2022) <https://www.bps.org.uk/news-and-policy/online-safety-bill-still-leaves-children-and-young-people-vulnerable-harm-and-must> accessed 1 July 2022.
[43] Online Safety Bill (As Amended on Report) s 201(2).
[44] ibid 201(3)(a); ibid 201(3)(b); ibid 201(3)(c).

essential to acknowledge that online safety regulations refer to young users of services as 'children' to distinguish them from adults. However, in literature and reports, this user group may be referred to as 'young people'. For accuracy, this paper will use the language originally used in reports.

The 2021 OFCOM Online Nation Report illustrated that 17% of young people surveyed saw content online that made them feel 'uncomfortable'[45], with 20% seeing something 'scary or troubling online like a scary video or comment'[46]. These qualitative descriptors of personal experiences could be interpreted as being equal to 'psychological harm'[47]

under both the provisions of the May 2021 and the December 2022 Draft Bills. In both iterations, the definition of harm could theoretically regulate and reduce the types of harm impacting young people. The explanatory notes accompanying the May 2021 Draft Bill indicate that the consistent terminology of 'psychological harm'[48] should include more serious harms than those transient, with 'longer-term conditions such as depression and stress; and medically recognised mental illnesses, both short-term and permanent.'[49] Indicating that in their deliberations of 'content harmful to children' platforms should be considering such a long-term impact, arguably making such challenging to adhere to. As the long-term impact of a singular piece of content is arguably challenging to identify, prompting a need for further guidance. Even when viewed in line with its wider legal origins of duties of care and tortious liability, the definition provided in such cases as *McLoughlin v O'Brien* of 'positive psychiatric illness'[50] provided by Lord Bridge, the reported experiences of young people are unlikely to meet such a threshold and therefore go unregulated. A new establishment of a context-appropriate definition and understanding of harm, clearly defined, would be beneficial as a forward step, as this would encompass reported experiences of harm, arguably laying the groundwork for effective harm reduction. Such a definition has been provided within Australia – a jurisdiction that has been developing its own similar online safety regulations in parallel – where harm has been considered within case law to occur where there is the presence of 'distress, alarm, fear, anxiety, annoyance, or despondency, without any resulting recognised psychiatric illness'[51]. Within the UK, the Protection from Harassment Act 1997 does provide for the awarding of damages where anxiety has been caused[52], introducing another source where such an understanding within the jurisdiction may be taken from. As well as another consideration for the end regulator as they current plans will see them impose a fine for a failure to adhere to duties, without a specific designation of types of harm content has caused. Such a specific list of recognised examples could benefit the UK's regulations and eventual real-time identification of content posing harm to children, as platforms would have identifiable examples to compare anticipated impact with. This is arguably dependent on the platforms' interpretation; it could be included in explanatory notes or subsequent codes of practice issued by the regulator, OFCOM, themselves.

Additionally, when juxtaposed with the potential shortfalls of the provisions, the Online Nation findings indicate that speaking with young people directly may help gather specific examples of online harm. This information can then be used to create more effective regulations that will benefit them by addressing the

[45] OFCOM, 'Online Nation Report 2021' (n 26) 79.
[46] ibid.
[47] Draft Online Safety Bill s 137; Online Safety Bill (As Amended on Report) s 201.
[48] Draft Online Safety Bill s 137; Online Safety Bill (As Amended on Report) s 201.
[49] 'Explanatory Notes to the Draft Online Safety Bill 2022' para 273.
[50] *McLoughlin v O'Brien* 432.
[51] *Tame v New South Wales* (2003). 211 CLR 317
[52] Protection from Harrassment Act 1997 s 3(2).

issues they are reporting instead of continuing with the vague terminology with unclear meanings set to regulate the current content they will see online.

The December 2022 Draft Bill sees the introduction of the threshold of a 'material risk of significant harm'[53], which must be met for a piece of content to be considered harmful to children. This widens the potential impact that platforms will have to consider, outside of the consideration of harm in this context, to designate a singular piece of content as harmful. Currently, larger user-to-user services such as Facebook, Instagram, and Twitter encourage user reporting and incorporate such into their regulatory frameworks. These reported instances, on top of findings published in reports like the Online Nation series, potentially provide a shortlist by which comparisons could be made. However, the historic lack of transparency reporting concerning the former has reportedly led to lessened use of mechanisms[54], indicating such a shortlist would be challenging to assemble. Whilst there has been some development in connection to harm, as discussed above, the newly introduced 'material risk'[55] is largely undefined. Again, looking at the tortious origins of these regulations could provide platforms with a starting point for recognising where content poses a risk to children and taking action to remove it. Within *Sienkiewicz v Greif (UK) Ltd*, Lord Phillips PSC held 'material'[56] to be more than de minimis[57], likewise in *Carder v The University of Exeter* 'material' was held to be 'more than negligible'[58]. Suggesting that for a piece of content to be determined as harmful to children, such would pose a level of risk to children that is more than minimal. On the one hand, this provides some general guidelines for where a threshold of 'material risk of harm' exists. However, this offline interpretation will need to be defined in context, giving rise to a need for further development.

As highlighted within the overview of changes, the December 2022 Draft Bill now appreciates how the context in which content is disseminated can give rise to harm. Section 201 states that harm to a user could occur where 'content [is] repeatedly sent to an individual by one person or by different people'[59]. This indicates that frequently reported harms such as cyberbullying must be considered and acted upon by platforms, an action welcomed when assessing these in line with the reported experiences. The 2021 Online Nation report illustrated that more than half of young people surveyed reported bullying online through direct messaging applications (54%) or social media platforms (53%) as an example of the harm they encounter online.[60] It can be theorised that the development showcased in the December 2022 Draft Bill has occurred due to policymakers to ensure that cyberbullying has reported harm is to be caught within the parameters of harm now presented, as this definition of harm forces platforms to take action to reduce content harmful to children in line with their Section 54 duties. However, this expansion could go further to provide a workable and effective definition in practice. The December 2022 Draft Bill states that content must be 'repeatedly'[61] sent to an individual. Yet, there is no further clarification as to whether this needs to be the same piece of content multiple times or the extent to which different pieces of content would have to be similar to be considered harmful. Nor is there an indicated time frame or limitation in which these repeated acts would need to occur to be recognised as harmful. This prompts further questions and a need

[53] Online Safety Bill (As Amended on Report) s 54(3)(c).
[54] OFCOM, 'Online Nation Report 2022' (OFCOM 2022) 72 <https://www.OFCOM.org.uk/__data/assets/pdf_file/0023/238361/online-nation-2022-report.pdf>.
[55] Online Safety Bill (As Amended on Report) s 54(3)(c).
[56] ibid.
[57] *Sienkiewicz v Greif (UK) Ltd* [2011]. UKSC 10. 40.
[58] *Carder v The University of Exeter*. EWCA Civ 790. 26.
[59] Online Safety Bill (As Amended on Report) s 201(3)(c).
[60] OFCOM, 'Online Nation Report 2021' (n 26) 80.
[61] Online Safety Bill (As Amended on Report) s 201(3)(c).

for clarification on the exact element that must be repeated for content to be considered a 'harm' under the newly drafted provisions. Again, there is an opportunity for such clarification from the regulator, OFCOM.

The reliance on 'content' within this expansion so raises concerns. As the December 2022 Draft Bill provides that content is 'anything communicated by means of an internet service, whether publicly or privately, including written material or messages, oral communications, photographs, videos, visual images, music and data of any description;'[62] Encompassing a wide range of content that could be posted by users, which in theory covers a comprehensive way in which harm could be caused. However, such a definitive list arguably presents challenges for the long-term regulation of these spaces where new methods of communication are evolving. Within the May 2021 Draft Bill, user-generated content was stipulated to be content 'generated by a user of the service, or uploaded to or shared on the service by a user of the service'[63] This is arguably a more workable definition when the day-to-day implementation of provisions is considered as such a broad definition would allow for the development of new methods communication to be designated as harmful, such as the emoji. Emojis have been previously raised in connection to both cyberbullying and racist comments in online safety committee proceedings[64], where witnesses raised doubt as to the capability of automated moderation systems utilised by larger platforms to regulate their services to recognise the harmful and discriminatory use of such emoticons and memes. To enact the closed list definition present in the December 2022 Draft Bill is to interpret that such is covered under 'written material'[65] or 'data of any description'[66]; however, without clarification from the regulator, this is a supposition and would need to be confirmed for definite when the long-term workability of these provisions is considered.

Theoretically, these developments – expanding the definition and including context – should benefit young people and children experiencing harm due to their online interactions. However, it is also possible that these provisions, if enacted as worded in the December 2022 Draft Bill, will place a regulatory burden so significant on platforms that such becomes unworkable in practice, thus causing no or little change to the level of harm experienced.

The introduction of an 'appreciable number' – determining an unknown user group.

Throughout their development, particular attention has been given to children as a user group within the online safety regulations. Both the May 2021 Draft Bill and December 2022 Draft Bill maintain exclusive provisions focused on the regulation of online services for the benefit of children and denote therein descriptors of the group; it is here where they differ.

Previously, the May 2021 Draft Bill denoted that for a service to be obligated to act against a piece of content on the basis that it could be harmful, such would have to have an impact on a 'child of ordinary sensibilities'[67] indicating that impact should be measured against a hypothetical individual. Within the December 2022 Draft Bill, this individual has been replaced with a group. Section 54 states that impact must be measured against 'an appreciable number of children in the United Kingdom.'[68] This move towards

[62] ibid 203(1).
[63] Draft Online Safety Bill s 39(3).
[64] 'Joint Committee on the Draft Online Safety Bill - Evidence Session No. 1' 27.
[65] Online Safety Bill (As Amended on Report) s 203(1).
[66] ibid.
[67] Draft Online Safety Bill s 45(3).
[68] Online Safety Bill (As Amended on Report) s 54(4)(c).

a threshold of many users rather than an individual arguably complicates the provision, with such potentially leading to content being excluded from regulation as such could be designated as not harmful.

The move away from an objective standard is welcomed when the potentially diverse characteristics of young people are considered, with 'an appreciable number'[69] now encompassing all young people rather than those who fit the previous unclarified descriptor of being of 'ordinary sensibilities'[70] widening the number of children platforms are obliged to consider when determining the presence of harm. Whilst a notably important change when inclusivity is considered, this removal then emphasises the 'appreciable number' [71] to encompass those groups that may have been excluded under the hypothetical standard of 'child of ordinary sensibilities'[72]. Although, if the 'appreciable'[73] threshold functions as currently envisioned, this will see the inclusion of all, which could be unworkable in reality, meaning that smaller categories of children harmed by certain content – for example, the notable story submitted in evidence at the Committee Stage concerning children with epilepsy being particularly harmed by videos of flashing lights - previously excluded under the sensibilities provision remain so due to the regulatory burned of recognising all being placed on platforms. On the other hand, this amendment is a move away from the akin legal standards of care, with Smith[74] previously suggesting that the inclusion of the 'child of ordinary sensibilities'[75] as a character of legal fiction was an attempt to align the new digital duties with those historically established in the common law. This indicates that this move towards a more qualitative standard is untethering the regime from the other duties of care present within England and Wales, meaning there is less baseline information for platforms and eventual regulators to look to for legal guidance.

Whilst welcomed and more encompassing of children of differing characteristics, it can be suggested that the alterations to the descriptions of 'children' when determining impact are neither an improvement nor deterioration of the section's overall workability. It can be argued that both 'a child of ordinary sensibilities'[76] and 'an appreciable number'[77] present unquantifiable standards and thresholds concerning any impact determined. Both present standards that are not expanded on or developed in either the Bill's or accompanying explanatory notes, leaving platforms blind to the user group they should have in mind when contemplating the impact of content to determine the presence of harm. The move away from the objective standard presented in the May 2021 Draft Bill questions how a company or platform should recognise and conceptualise the largely unquantifiable standard of 'an appreciable number'[78] concerning each piece of content on their platforms that could harm users.

Companies or platforms need to adequately identify and conceptualise such a demographic to avoid harmful content being left on platforms, putting the publicised goal of the Online Safety regime at risk. With the current wording, content may be either overregulated or not regulated at all, each of which has consequences for the safety and freedoms of young people online. On the one hand, without a recognisable threshold or descriptor to have in mind by which any potential impact could be mapped or assessed, a piece of content

[69] ibid.
[70] Draft Online Safety Bill s 45(3).
[71] Online Safety Bill (As Amended on Report) s 54(4)(c).
[72] Draft Online Safety Bill s 45(3).
[73] Online Safety Bill (As Amended on Report) s 54(4)(c).
[74] Graham Smith, 'On the Trail of the Person of Ordinary Sensibilities' (*Cyberleagle*, 29 June 2021) <https://www.cyberleagle.com/2021/06/on-trail-of-person-of-ordinary.html>.
[75] Draft Online Safety Bill s 45(3).
[76] ibid.
[77] Online Safety Bill (As Amended on Report) s 54(4)(c).
[78] ibid.

that could cause harm to users could be left on a platform unregulated. On the other hand, content that does not cause harm could be taken down due to platforms and services misinterpreting such regulations. More likely, they will be incentivised to be over-cautious and remove content that does not pose a risk of harm when combined with this broad and vague terminology adopted, becomes even more likely.

Based on the current system of moderation present on many services, the identification and takedown of this content will likely be carried out by automated detection systems or algorithms. The standardised idea of harm presented in the December 2022 Draft Bill is likely to be challenging to build into these algorithms, given the need for more publicly available information about the computational elements of systems.[79] This leaves two potential outcomes: an overburdened, ineffective regulatory arm of services reliant on human processing lacking complete oversight due to an inability to consider every aspect of service and every eventuality of harm or a plethora of harmful content remaining on platforms due to the lack of automated detection systems being utilised. Neither supports the aim of the provisions to reduce the risk of harm.

The Relationship between Harmful Content and the Definition of 'Harm'

As has been determined within this paper, harm is a concept central to the online safety regulations, being frequently referred to within both the December 2022 Draft Bill and associated documents. Not only is the specific definition of harm unclear as previously established, the link between Section 54 – outlining content considered harmful to children – and Section 201 – the definition of harm – is unclear. This lack of clarity, alongside that present within the sections themselves, is apparent in the overall layout of the provisions.

Previously within the May 2021 Draft Bill, content considered harmful to children would be determined harmful if such had the risk of a 'significant adverse physical or psychological impact'[80] to a child user, whereas the December 2022 Draft Bill lacks the inclusion of 'impact'[81] instead favouring 'material risk'[82]. The move away from 'impact'[83] on a user in favour of 'risk'[84] to a user arguably places emphasis and expectation on the definition of harm – as previously argued, this needs further development – and thus the connection between the two provisions to be clear. Previously under the May 2021 Draft Bill, a platform could look at the context in which the provision operates to understand what 'impact'[85] is alluding to, content that will affect children in some way. Whilst the development of the provision has seen the title remain the same, the removal of 'impact'[86] leaves platforms lacking context and understanding as to what material risk could be, as previously explored. Therefore, the link between the section which defines harm – Section 201 within the December 2022 Draft Bill – and that which outlines content harmful to children – Section 54 – becomes even more important. Yet, such a link is absent in its entirety.

The absence of specific reference to Section 201 in Section 54, and vice versa, is concerning when the overall intelligibility and workability of the provisions are considered. The complex layout of the Draft legislation that was published in March 2022 has been previously commented on, with Perrin et al.

[79] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015).
[80] Draft Online Safety Bill s 45.
[81] ibid.
[82] Online Safety Bill (As Amended on Report) s 54(4)(c).
[83] Draft Online Safety Bill s 45.
[84] Online Safety Bill (As Amended on Report) s 54(4)(c).
[85] Draft Online Safety Bill s 45.
[86] ibid.

suggesting that the structure is 'too difficult to navigate'[87] with crucial elements such as definitions being 'buried at the back, rather than upfront'[88], indicating that principal phenomena such as what is meant by harm and where this is essential to understanding key sections, should be clearly defined and referred to throughout. Whilst the layout of this legislation is not uncommon – with the interpretations and explanatory sections being placed in the back as opposed to their accompanying sections throughout – the structure for a set of provisions with a flagship term such as 'harm' should arguably be more explicit. The term should be defined before any in-depth detailing of specific duties and obligations relying on the universal definition. This current layout, without an upfront definition, has been suggested to be 'cumbersome'[89] for parties who may not know to search all the seemingly unconnected regulations for definitions related to successfully discharging their duties. This is unlikely where the platforms discharging their duties are larger companies such as Facebook or Twitter, which have larger legal departments to assess and comprehend the regulations and know to search the entire Bill for definitions related to the successful discharging of their duties. However, as these duties relate to 'services likely to be accessed by children', it is possible that platforms regulated will encapsulate smaller platforms and small to medium enterprises with lesser or no legal departments. For this reason, the relationship between the provisions outlining content considered harmful and that providing the definition of harm must be clarified, as without reference being made to the broader meaning of harm within this context, it is possible that content could be left unidentified and present on platforms, leading to potential ineffective reduction of online harms.

Development of secondary legislation and guidance as a way forward?

This paper has highlighted potential issues and challenges that could arise should the December 2022 Draft Bill be enacted as currently written. It has also been illustrated that online safety provisions are progressing through the legislative process and are potentially subject to further changes and developments. Incumbent officials have theorised and commented on elements such as the designation of primary priority content and priority content as the Bill has moved through the legislative process[90], with further clarifications expected in time. However, elements crucial to regulation – such as the vague definition of harm and which commonplace communication methods are likely to be considered content – are left to further guidance and the discretion of the regulator, OFCOM.

This transitional nature of the provisions, as they progress through the Parliamentary process, is beneficial when considering potential resolutions and ways forward, as there is both more time for contemplation of the issues raised above and an opportunity to ensure that these are a priority for the secondary legislation expected after the enactment of the regulations.[91] OFCOM play a significant role in the overall intelligibility and effectiveness of the online safety regulations, developing and maintaining 'codes and guidance on protection of children'[92], which are anticipated to be finalised in late 2024[93] These codes and guidance are

[87] William Perrin, Lorna Woods and Maeve Walsh, 'The Online Safety Bill: Our Initial Analysis' (Carnegie Trust UK 2022) 1 <https://d1ssu070pg2v9i.cloudfront.net/pex/pex_carnegie2021/2022/03/31120201/The-Online-Safety-Bill-Our-Initial-Analysis.pdf>.
[88] ibid.
[89] ibid 3.
[90] Nadine Dorries, 'Online Safety Update Statement Made on 7 July 2022' (*UK Parliament*, 7 July 2022) <https://questions-statements.parliament.uk/written-statements/detail/2022-07-07/hcws194> accessed 18 July 2022.
[91] Andre Rhoden-Paul and Kate Whannel, 'Online Safety Bill Put on Hold until New Prime Minister in Place' *BBC News* (13 July 2022) <https://www.bbc.com/news/uk-62158287> accessed 18 July 2022.
[92] OFCOM, 'Online Safety Bill: Ofcom's Roadmap to Regulation' (OFCOM 2022) 7 <https://www.ofcom.org.uk/online-safety/information-for-industry/roadmap-to-regulation>.
[93] ibid 17.

expected to be crucial when the overall intelligibility and workability of the Bill overall is considered, as at present vital elements of the provisions are unclear with a need for further clarification. OFCOM's codes and guidance are expected to clarify what types of content could be deemed harmful and provide an opportunity for harms not yet envisioned to be regulated.

The eventual regulator, OFCOM, has been confirmed to provide oversight for enforcing the regulations, acting like the Information Commissioner's Office concerning data protection. As part of their role, OFCOM will further clarify matters such as the meaning of 'material risk'[94] in practice and how services can effectively identify this. The approach to providing pertinent information such as this was outlined in the Government's response to the Online Safety Committee's findings, where they stated that they 'still believe that setting out the priority harms to children in secondary legislation is the better approach'[95] confirming that secondary legislation and guidance – the anticipated 'codes of practice' – will be the primary vehicle by which further clarification on vague or unclear language and interpretations currently present. However, the determinations and guidance in these codes cannot be consulted on or planned until the secondary legislation defining priority and primary priority content is finalised. The primary legislation has been delayed by an estimated six months, suggesting that the more extended such clarification in secondary legislation takes, the longer potentially problematic and harmful content will remain on platforms, ineffectively regulated.

This reality of postponement is likely to be problematic for successfully reducing online harm. As mentioned previously, the Secretary of State incumbent in June 2022 indicated their plans for the content they would be determined as a primary priority and priority content providing the start of a workable list for regulation. As illustrated above, this will benefit all stakeholders in managing regulation expectations and provide clear examples of what content should be regulated urgently. However, this Bill is arguably related to the broader political aims and agenda and, therefore, could align with the political landscape. The Bill itself could yet undergo further changes until it is enacted. Despite the potential uncertainty surrounding the Bill as part of the Online Safety regime, OFCOM is set to remain as the regulator as indicated in the Bill's long title: 'A Bill to make provision for and in connection with the regulation by OFCOM of certain internet services; for and in connection with communications offences; and for connected purposes.'[96] This indicates that OFCOM can be identified as a certainty in relation to the Bill and, therefore, could be called upon to provide further clarification and guidance to service providers. As a regulatory body, they are most equipped to bring into force quick changes. They would successfully support the Bill's plans to leave determinations to secondary legislation. This plan is perhaps best for protecting children, given the rapid turnaround of online trends and platform engagement. Regulatory codes and guidance will have the potential to utilise phenomena such as previous judicial decisions and academic findings to adapt that best follows the main provisions to protect young people and reduce harm online.

A potential avenue to resolving these issues is utilising reports such as the Online Nation series featured here to build guidance notes and codes of practice that directly tackle instances of harm reported by young people. So far, discussions concerning these regulations have focused on and stressed the importance of protecting children as a user group of online services, with minimal mention of engaging with these and those who represent them directly as part of developing secondary legislation. Such engagement is possible with charities representing children, such as 5Rights and the Molly Rose Foundation, invited to give written

[94] Online Safety Bill (As Amended on Report) s 54(4)(c).
[95] Department for Digital, Culture Media & Sport, 'Government Response to the Report of the Joint Committee on the Draft Online Safety Bill' (2022) 34.
[96] UK Parliament, 'Online Safety Bill - Parliamentary Bills - UK Parliament' (*UK Parliament*, 2022) <https://bills.parliament.uk/bills/3137> accessed 22 August 2022.

and oral evidence to the Online Safety Bill Committee in September 2021. These were opportunities for the harms impacting children and young people to be showcased, reflected upon, and addressed within the Bill. Some of these changes between the May 2021 Draft Bill and the December 2022 Draft Bill are a direct consequence of such engagement. Such as appreciating the context in which harm could occur now potentially encompassing cyberbullying, previously difficult harm to regulate. However, this also shows that such engagement is a beneficial route forward in connection to the primary legislation and the secondary legislation and guidance as a subset of regulations. It is noted that providing regulations for such a wide-ranging user group with distinct requirements beyond their age is challenging. Therefore, it is vital that such baseline terms as 'harm', 'material risk'[97], and 'appreciable number'[98] are clarified in advance to create a workable framework which can be built on via findings of direct engagement.

Conclusions and Further Works

It is pertinent to note that due to the above-discussed ongoing nature of the development of these regulations, the conclusions reached within this paper indicate areas where further work and developments need to be considered. This paper has shown that the changes between the May 2021 Draft Bill and December 2022 Draft Bill will potentially leave harmful content online, where it can pose a risk to children. It has been illustrated that some of these alterations and removal of specific wordings entirely, whilst welcomed for the overall development of the regulations, could have ramifications for the overall aims and effectiveness, potentially failing to reduce the amount of harmful content online.

Overall, there is an urgent need for clarification and further guidance in connection to critical terms such as harm and 'material risk'[99], as they are presently superficially defined to the extent that they will not provide the level of guidance that platforms will need to adhere to their duties and reduce harmful content successfully. Whilst the potential issues and challenges have been highlighted in connection to the latest published iteration of the provisions, the changes and developments occurring between the publication of the May 2021 Draft Bill and the December 2022 Draft Bill do indicate that there is potential for further development in response to such criticism, leaving the possible future of secondary legislation hopeful.

It has been exhibited that to be fully effective at identifying and reducing harms online, there will need to be further guidance concerning two vital provisions, which outline the content that should be determined as harmful to children and that which defines harm. The conclusion of the December 2022 Draft Bill being potentially ineffective in practice if enacted as currently worded is tentative in nature. This iteration of the provisions is now going through the latter stages of legislative development. Therefore, this paper has presented a solution that looks towards the secondary legislation and the challenges that could occur here. This has been publicised as that which will 'tidy up' any unclear parts of the regulation. However, to rely on the strength and rigour of secondary legislation where such is typically expected of primary legislation is perhaps a gamble on the part of those who have envisioned the UK as a place online where there is a reduced level of harm in comparison to the rest of the world. Whilst OFCOM does seem well equipped to tackle such challenges, there is a need for further work to ensure that these provisions reduce harm for target user groups in their everyday interactions online.

This paper has shown that the currently worded provisions have two main issues that must be addressed. These issues have been identified, and it is expected that resolving them may take some time, presenting opportunities for further work beyond the drafting of clearer codes and guidance. Firstly, there is a need for

[97] Online Safety Bill (As Amended on Report) s 54(c).
[98] ibid.
[99] ibid 54(4)(c).

direct engagement with the user group of young people to gauge their opinions and experiences with platforms and online harms to determine their understanding and views on the suitability of the Bill introduced. This will allow for conclusions to be reached as to the (or any) potential effectiveness of the Bill's provisions for the reduction of online harm.

Secondly, whilst there is some debate around the next steps for OFCOM as the regulator, this paper has shown and affirmed that secondary legislation is the appropriate legal vehicle for clarification. Therefore, it is logical for there to be further work within the area to explore and understand how secondary legislation will operate within this sphere. Such will be particularly pertinent when the multistakeholder, cross-jurisdictional nature of the internet and the platforms that operate within such are considered. Therefore, further work combined with direct engagement will likely assist the development and eventual operation of the regulations in delivering the promise of harm reduction online.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council EP/S023305/1