

Treating the symptoms or the disease? Analysing the UK Online Safety Act's approach to digital regulation

Victoria Nash  | Lisa Felton

Oxford Internet Institute, University of Oxford,
Oxford, UK

Correspondence

Victoria Nash, Oxford Internet Institute,
University of Oxford, Oxford, UK.
Email: victoria.nash@oii.ox.ac.uk

Abstract

In recent years, the pace of Internet regulation around the world has quickened, with states increasingly confident that they can and should hold major platform companies to account. New laws have been developed to address the risks of digital technologies and law-makers have drawn on familiar regulatory principles and legacy frameworks in addressing them. But the nature of the technologies and the business models supporting them bring new challenges which make it less clear that old approaches will work. To succeed, legislative frameworks must evolve and adapt. Against this backdrop we assess the UK's Online Safety Act 2023 (OSA), which was expected to provide an innovative and broad-reaching 'systems-based' approach to reducing user risks and harms, particularly in relation to child safety. We argue that although the Act does incorporate measures to regulate platform design, it fails to fully embrace this and faces some challenges in ensuring proportionality and accountability. We conclude that the development of the OSA is hampered by a legacy focus on content controls which may limit its ability to effectively improve online safety, particularly as services evolve.

KEYWORDS

digital policy, Internet regulation, Internet safety

Research conducted during a Visiting Policy Fellowship at the OII, now completed. Current affiliation at the Vodafone Foundation.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Policy & Internet* published by Wiley Periodicals LLC on behalf of Policy Studies Organization.

INTRODUCTION

In recent years, the pace of Internet regulation has quickened around the world, with states increasingly confident that they can and should hold major technology companies to account through democratically scrutinised legislation. Such legislation has addressed a wide range of policy challenges, including market dominance and data monopoly (EU Digital Markets Act 2022), privacy and data protection (EU General Data Protection Regulation 2018 and California's Age-Appropriate Design Code Act 2022), AI governance (Canada's forthcoming Artificial Intelligence and Data Act and the EU's forthcoming Artificial Intelligence Act) and consumer harms (Australia's Online Safety Act 2021 and EU Digital Services Act [DSA] 2022). The latter category has garnered particular media and public attention against a backdrop of dissatisfaction with the perceived risks posed by social media and other services in an economy which monetises personal data by exploiting human attention.

Some of these challenges are familiar, and as such, law-makers have understandably drawn on familiar regulatory principles and legacy frameworks in addressing them. But at the same time, the nature of the technologies and the business models supporting them bring new challenges which make it less clear that old approaches will work. To succeed, legislative frameworks must evolve and adapt. Against this backdrop, we assess the UK's Online Safety Act (OSA), which was expected to provide an innovative and broad-reaching systems-based response to concerns about user risks and harms.

LITERATURE REVIEW

From an academic perspective, the continuing evolution of Internet regulation through instruments such as the OSA takes place against the backdrop of a long-running debate about the nature and responsibilities of online service providers (Flew et al., 2019; Gillespie, 2018; Napoli & Caplan, 2017). Traced back to the structural distinction drawn in US media policy between carriers and content (Flew et al., 2019), this debate asks whether such services are best understood as publishers of content, with accompanying responsibilities for the content provided, or as 'dumb pipes', carrying material provided by third parties with no or limited responsibility for this material. As Kosseff, (2019) noted, the latter view, instantiated within the Communication Decency Act's Section 230, undoubtedly played a major role in energising Internet innovation. The CDA held that Internet service providers would not be held responsible for illegal content hosted on a site, and would retain this immunity even when choosing to edit or remove content. The EU equivalent, the 2002 e-Commerce Directive included similar but less extensive safe harbour conditions, offering online services conditional immunity only up to the point where they become aware of illegal content hosted on their sites (Flew, 2021).

The effect of these dual regimes has been to crystallise the view that online hosting services are not content publishers. However, with the advent of social media giants such as Facebook, YouTube, TikTok and Instagram, and an increasing array of harmful and illegal content available, this view is subject to challenge (Flew, 2021). The majority of content on these services may indeed be provided by third parties, but the platforms play an expanding role in how that content is served to users. Content is curated algorithmically to attract our attention (Gillespie, 2018), is served via highly-designed platforms which encourage us to continually swipe or scroll to maximise time spent (Williams, 2018), is monetised through the sale of advertising using user data profiles (Zuboff, 2019), and is rendered more or less visible depending on concerns about its truth or potential for harm (Myers West, 2018). Against this backdrop, critics ask whether platforms can legitimately claim ignorance of the content posted (Flew et al., 2019), or be understood to play a neutral, rather than active role

in serving that content (Buiten, 2021). Here the difficulty lies in the bifurcated policy options proposed to date: is it still helpful to view social media platforms as necessarily either dumb pipes or publishers, or as neutral intermediaries rather than 'active co-creators of the digital sphere' (Buiten, 2021, p. 361)?

Potentially there are other options. Some have argued for a regulatory approach that tackles market dominance and promotes better consumer outcomes via enhanced competition (Fukuyama & Grotto, 2020; Helberger, 2020). This may help but is a blunt instrument that couldn't directly hold platforms responsible for the full range of harms. Another approach is to reconsider the distinction between infrastructure and content. In their paper for the Carnegie UK Trust, Lorna Woods and William Perrin argued that the problem lies with a regulatory focus driven solely by content-related concerns: 'the starting point is wrong' (Perrin & Woods, 2019, p. 11). Instead they propose a system-level approach that addresses platform design and architecture. Drawing on Larry Lessig's work, they argue that 'the environment within which harm occurs is defined by code that the service providers have actively chosen to deploy' (ibid p. 12), and as such, responsibility lies with providers for design choices which exacerbate or mitigate risk and harm. Platforms should be held responsible not for individual pieces of content hosted, on this view, but for exercising a duty of care in designing the systems in which such content is elicited, hosted, shared, promoted and monetised. Price (2021) takes a similar approach, but argues this is merited because social media function not just as a platform for content but a space where people congregate and amplify individual voices. Such a systems-level approach is not completely new, indeed, in addition to similarities with regulation governing provision of safe public spaces or work environments (ibid), it shares features with principles of security-by-design or privacy-by-design found in other areas of regulation (Cavoukian, 2009).

There are many possible benefits of a systems-based approach to regulating online platforms. Perhaps most simply, it aligns responsibilities. Under a systems-focused approach, platform companies are held responsible for the decisions they make (the systems they design and build), rather than the decisions users make (which content to post) (Demos, 2022; Price, 2021). As Internet studies scholars have pointed out, platforms already govern users' behaviour through such decisions (Gillespie, 2015; Gorwa, 2019; Plantin et al., 2018), so finding ways to ensure that these impactful processes accord to the rule of law is already normatively compelling (Suzor, 2018), while doing so through explicit government regulation provides greater transparency and democratic accountability (De Gregorio, 2022).

Additionally, measures to address problematic or illegal content are typically implemented only after material is shared, unless platforms apply automatic upload filters across huge swathes of content, a form of censorship limited under EU law (Rojszcsak, 2022). Such an ex post approach lends itself to measurement through metrics such as speed of content take-downs, with platforms then reporting these to demonstrate supposed efficacy of algorithmic moderation (Gorwa et al., 2020; Theil, 2019). In contrast, systems regulation applies to processes and design features across all stages of the user experience, but must necessarily be established ex ante. This approach is less likely to create perverse incentives for companies to prioritise speed or scale over accuracy. Another key feature of a systems-based approach is that it can be adjusted to take account of a wide variety of risks. As child Internet safety literature has long argued, individuals may experience harms from online contact, they may themselves cause harm to others through their own conduct and may be at risk in commercial relationships or contracts (Livingstone & Stoilova, 2021) (Table 1).

On the other hand, it's also important to understand the limitations of systems-based approaches. Pathways to harm are complex, arise in specific socio-technical contexts, with some individuals more vulnerable than others on the basis of personal and contextual factors. Although such measures have been portrayed as 'techno-legal

TABLE 1 Summary of differences between content and systems-based approaches.

Content measures	Systems measures
Addresses outcomes of the system.	Addresses risks of the system and its design features.
Typically applied ex post facto to address or remove content once posted.	Applied ex ante, preventing problems arising or mitigating risks that can't be prevented.
Metrics might include number of take-downs, speed of take-down or accuracy of decisions.	No easy metrics but targets governance, decision-making and process design, including remedies.
May create possible perverse incentives to meet regulatory targets in ways that limit freedom of expression or ignore harms.	Intended to address the perverse incentives that encourage engagement with risky content.
Company taking responsibility for users' actions.	Company taking responsibility for company's decisions.
Solely focused on risky content.	Can address a wide array of risks.
Grounded in legacy frameworks of content regulation	Flexible and future-proof.

solutionism' (Angel & boyd, 2024), systems-based responsibilities for platform designers should be understood as just one aspect of necessarily broader societal efforts to ensure positive digital experiences for children and young people. Education, parental support and measures that build digital resilience are vital, while also recognising that public policy need to address underlying problems such as poverty, insecurity and under-provision of child and adolescent mental health services.

An over-arching duty of care is not the only possible manifestation of a systemic approach to regulation. The fundamental requirement is simply that regulation should target platform design and "how technologies are...designed and deployed within online services" (Demos, 2022, para. 4). Jack Balkin's (2020) proposal that platforms should bear 'fiduciary' duties to protect users' interests is another well-developed approach. But insofar as this 'duty of care' approach was embraced by the UK government and elements are visible in the OSA, in the remainder of this article we ask whether the OSA marks a step-change in Internet regulation by embracing a systems-based approach, and if so, how likely this is to effectively tackle risks and harms for children.

THE UK OSA

The OSA has a long history, evolving through two rounds of public consultation in 2017 and 2019 and more than 2 years of parliamentary scrutiny and debate. The second consultation saw publication of the Online Harms White Paper (OHWP), which marked a significant advance for UK Internet policy, arguing that regulation was needed to tackle an extensive array of online harms, including illegal content, 'legal but harmful' content and other non-content-related harms such as over-use of devices (Nash, 2019). Crucially, the OHWP incorporated Perrin and Woods' (2019) systems-focused proposal, namely that in-scope companies should bear a preventative 'duty of care' to users in the design and delivery of online services. The resulting Online Safety Bill (OSB) was introduced into the UK House of Commons on 17 March 2022 and finally received Royal Assent on 26 October 2023.

The final Act's scope is very broad. The range of companies covered includes 'user-to-user' services and 'search services', with the largest and/or riskiest of the user-to-user

services designated as ‘Category 1’ services, subject to additional responsibilities, such as providing user empowerment tools for adults and protecting journalistic content. The Act imposes responsibilities to prevent user harm based on risk assessments of those services. All in-scope companies must conduct risk assessments for illegal content, and all also bear responsibilities to preserve freedom of expression and privacy. Services that are likely to be accessed by children face additional responsibilities and must conduct risk assessments and then act to protect young users from harmful content.

The Act has evolved significantly since the publication of the OHWP. The change of name, from ‘online harms’ to ‘online safety’ may not be hugely consequential but perhaps reflects the shift from historic approaches which sought to tackle harms *ex post facto* to a focus on ensuring safety by identifying and reducing risk. The duty of care concept is still explicitly included but isn’t presented as an over-arching guiding principle, with Part 3 of the Act devoted to duties of care, and parts 4, 5 & 6 devoted respectively to ‘further duties’, duties relating to providers of pornographic content, and payment of fees. The regulator charged with oversight is Ofcom, the UK’s communications regulator, founded in 2003 to regulate communications infrastructure. New sections were added after committee scrutiny, most notably one requiring providers of pornographic content to prevent access by minors and another establishing new communications offences around harmful, false and threatening communications and sending images or videos of genitals. Provisions to regulate ‘legal but harmful’ content for adults were notably scaled down as a result of party opposition, whilst measures to hold senior executives personally liable were added to prevent backbench revolt. Such alterations are a vivid reminder that UK digital policy is shaped not just by industry lobbying, but by internal and cross-party political power struggles.

Insofar as the focus on child safety has been the most consistent aspect of the Act’s evolution, the rest of this article will focus on this aspect of the legislation. Analysis is based on the Bill drafts, written evidence submitted to the Public Bills Committee and the final Act’s text, as well as emerging implementation information published by the regulator.

A TRULY SYSTEMS-FOCUSED APPROACH?

The White Paper that preceded the OSA introduced the concept of a statutory duty of care, based on a policy paper published by the Carnegie UK Trust (Perrin & Woods, 2019). Such an approach was intended to be preventative, risk-based and outcome focused, and was grounded in an understanding that online platforms do not merely host content or interaction, rather they shape it through their design choices. The key normative foundation of the OHWP was thus a focus on minimising harms by governing online platforms as systems, rather than just as hosts of content. Although this foundation is still present in the final legislation, it seems to competes with a more traditional, broadcasting-model focus on content risks.

One of the original authors of the duty of care framework noted early on that even the OHWP failed to adopt the full extent of their recommendations, querying whether there was enough focus on the ‘constitutive impact of service providers’ in creating risk, rather than just removing problematic content once identified (Woods, 2019, p. 15). This gap widened further with the move from the OHWP to the Online Safety Bill, in which the duty of care was presented as one of several duties rather than an over-arching framing principle. The gap was also evident throughout the text of Bill drafts, with companies in scope repeatedly expected to consider both the implications of content and platform functionalities. In their written evidence to the Public Bills Committee, the think tank Demos observed that: “The Bill – Links (*sic*) systems and safety duties to categories of content and not (for instance), types of harm as they arise from systems in the first place.” (Demos, 2022) and went on to argue that the Bill prioritises oversight of systems for content moderation and curation above other system-based sources of risk.

The resulting Act continues to list content risks separately, despite the fact that these risks and harms arise as a result of online services' systems and design, and could thus be covered by this. For example, the Act's section 12(2) requires companies to ensure their services are appropriately designed to:

"(a)mitigate and manage the risks of harm to children in different age groups, as identified in the most recent children's risk assessment of the service (see section 11(6)(g))", and

"(b)mitigate the impact of harm to children in different age groups presented by content that is harmful to children present on the service." (OSA 2023, s. 12(2))

Similarly, the Act's section 11(6) sets out specific duties on companies to conduct risk assessments, with eight subsections detailing matters which such assessments must cover. Three of those focus specifically on harms associated with content, whilst three cover functionalities, algorithms, and the design and operation of the service. On the latter front, risk assessments must consider how the design and operation of the service (including the business model, governance, use of proactive technology and measures to promote users' media literacy) may reduce or increase the risks identified (OSA S.11 (6)(h)).

Interestingly, Hansard, the official parliamentary record, reveals that some of the clauses on system design were only included as a result of amendments brought by the House of Lords. The government pushed back on any further emphasis on functionality and design, citing concerns for "legal confusion" and delay (Hansard Online Safety Bill 12 September 2023). One amendment was agreed and Section 11 (6)(f) of the finalised Act thus states that a children's risk assessment should take into account

"... the different ways in which the service is used, including functionalities or other features of the service that affect how much children use the service (for example, a feature that enables content to play automatically), and the impact of such use on the level of risk of harm that might be suffered by children." (OSA 2023s.11(6))

That this addition was even needed demonstrates how far the OSA diverges from a more consistently systems-focused approach to retain a fascination with content-based risk.

To be clear, it is not that content harms don't matter. Rather, as noted above, content-related harms are just one type of harm that might arise from design decisions taken by platforms, and it seems unsatisfactory to privilege these above others, even if prior international regulation has often focused on content (Price, 2021). A more coherent and consistent system-based approach would involve requiring companies first and foremost to consider the wide range of risks and harms arising from their design choices, not dissimilar from the Australian eSafety Commissioner's recommendations regarding 'safety by design' (eSafety Commissioner Australian Government, 2019). This would also have the advantage of being a potentially more future-proof approach—content may not be the most significant source of harm in the metaverse for example (Demos, 2022), while the risks of generative AI tools if incorporated into search and social media will likely extend far beyond the content they serve (Weidinger et al., 2021).

FEASIBILITY OF THE MEASURES PROPOSED

Insofar as the OSA incorporates attempts a systems-based approach, albeit imperfectly, the second half of our question concerns the feasibility of the measures proposed. On the latter point, time will obviously tell, but there are three areas of weakness which may undermine

the Act's efficacy: proportionality, accountability, and the vagueness of measures surrounding age verification. These will be addressed in turn.

Proportionality: A tiered approach versus obligations for all

One of the challenges of applying a statutory duty of care to online services is the difficulty of determining what constitutes *reasonable* care. How much should companies invest in ensuring their services are safe to use? Companies of different sizes and at different stages of maturity will have varying levels of resource and expertise to apply to the problem, while technologies are constantly evolving and the associated risks will likely affect different user groups in different ways and may be hard to predict (Slavcheva-Petkova et al., 2015).

Proportionality can be achieved in a variety of ways, for example, by applying different rules to different companies based on size and resources, specific content or nature of the user base (e.g., children). Regulators should also provide clear guidance on what is required, reducing costs for businesses and ensuring clarity. Perrin and Woods (2019) recommended guidance be provided on risks and how to mitigate these using simple techniques such as decision trees, libraries of 'good code' which address common risks, and training and development, especially for start-ups and SMEs (ibid, p. 47), a point reiterated by the children's digital rights NGO 5Rights who argue that smaller providers need "support to comply, not permission to harm" (5Rights Foundation, 2021, p. 5).

Taken at face value, the OSA builds in limited consideration for proportionality. Section 13 (1) does state that to determine what will count as proportionate measures to ensure children's safety, relevant factors will include the nature and severity of the risk identified and the size and capacity of the service provider. But the Act requires that *all* providers of regulated user-to-user services likely to be accessed by children must comply with specific duties in relation to online safety. This applies to a huge range of companies, extending far beyond the largest and most familiar social media providers. Section 12(2) of the Act sets out a duty for these companies "to take or use proportionate measures relating to the design or operation of the service" to reduce and manage the risk of harm to children in different age groups, as identified in the most recent risk assessment. Thus each provider must complete a risk assessment accounting for the user base (including the number of users who are children in different age groups) and the level of risk. As noted above, risk assessments apply both to content and systems, taking into account algorithms used by the service, how quickly and easily content may be disseminated and functionalities presenting higher levels of risk.

The completion of risk assessments and a commitment to undertake risk-mitigating measures are thus duties that apply to all services likely to be accessed by children, regardless of their size. But does this mean that a start-up has to invest the same resources as a major international platform? Other regulatory bodies have been careful to embed considerations for proportionality in their approach to law-making. The Australian Online Safety Act (2022) requires industry bodies to develop codes which establish system level obligations across different services, with those obligations reflecting relative levels of risk. The EU's DSA applies different rules to intermediaries based on their size, role and impact. The tiered requirements start with a baseline applicable to all intermediaries including transparency reporting, cooperation with national authorities and points of contact. At the next level, additional requirements apply to hosting services, online platforms and very large online platforms and search engines (VLOPs and VLOSEs), including notice and action and reporting criminal offences. The next tier of obligations apply only to online platforms (including VLOPs), and include a wide range of obligations, including complaint and redress mechanisms, trusted flaggers, bans on targeted ads for children and those based on special

characteristics of users. Finally, a set of obligations apply solely to the VLOPs and VLOSEs, including risk management obligations, external & independent auditing requirements, and an internal compliance function.

Although the OSA text does not incorporate such a tiered approach as a means of ensuring that companies face proportionate regulatory duties, there is some evidence that the regulator, OfCom, will, in practice, hold different companies to different standards of care. The Code of Practice documents relating to implementation of the Act's child safety measures have yet to be published at the time of writing, but draft guidance covering illegal harms have been issued (at this date, for consultation).¹ In this draft guidance, proportionality is considered in terms of the obligation, level of risk and size of provider. In addition, the guidance does provide examples of recommended mitigation measures include the implementation of an *annual* review of risks and also reviews to identify new kinds of illegal content, both of which should also apply in guidance for services accessed by children. The draft guidance also suggests a number of systems-based measures which could be incorporated to protect children from illegal harms, including limiting their visibility on the service, restricting the ability to send messages from non-connected accounts, hiding their location, and provision of enhanced user control.²

Incorporating such measures in regulatory guidance rather than primary legislation is an interesting approach. It allows flexibility and agility in the face of rapidly-advancing technological change, but also places great power in the hands of the regulator.

Accountability and oversight

A second challenge of a systems-based approach concerns accountability. Under more traditional content-based regimes, publishers and broadcasters could be held to account for illegal content disseminated via their services. Assessments of liability could be made '*ex post*', with clear evidence of failure to abide by the law. But a systems-based such as that proposed by Perrin and Woods relies on assessing the adequacy of preventative measures—an *ex ante* approach. Given that the OSA requires in-scope online services to exercise a duty of care, how can the UK regulator assess this effectively, ensuring these companies are held to account for their actions?

In practice, even the findings of company risk assessments may vary widely. Some providers may consider their existing tools and protections to be sufficient, whereas others may implement a more rigorous and comprehensive review across their entire service. This is possible because the OSA allows companies to use their own subjective interpretation of risk as it arises on their services. As noted above, potential online harms are complicated, may vary between services and may not affect all users equally, whilst specific features may offer both opportunities and risks. This leaves scope for considerable variation in company responses to risk, exacerbated by the fact that the requirements apply to all services accessible to children, both large and small, meaning that the guidelines are likely, at best, to set a minimum baseline.

The challenge involved in assessing online services' efficacy in managing user risks is evident in experience of earlier regulation. The UK's regulation of video-sharing platforms (VSPs), introduced in 2020, is a useful point of comparison. According to the Communications Act 2003, VSP providers are required to comply with notification requirements and protect users from harmful content, with additional protection for children. The Ofcom Guidance for service provider recommends risk assessments (Ofcom, 2021). Following the first year of implementation, Ofcom published a report on the regulation's effectiveness at a point where 19 VSPs had submitted their notification (Ofcom, 2022a). Three key issues were highlighted which are potentially relevant to the OSA. The first was that platforms provided limited evidence on the effectiveness of their safety mechanisms.

Second, access control mechanisms were seen to be ineffective, giving rise to risks particularly in relation to access to pornography. And thirdly, platform providers were not prioritising risk assessments. Even with these more targeted and specific obligations, OfCom's own report indicates that a unilateral approach is flawed unless supported with additional accountability requirements, such as transparency reports, fully independent auditing, access to data for researchers (critical as the number of services caught in the regulation increases), guidelines and training. The OSA captures a much wider cohort of services within the remit of its regulation, putting a huge burden on the regulator and raising the question of how effective oversight will be possible.

This is another area where the DSA takes a different route. In addition to taking the size and scale of platforms into account when setting out new responsibilities, the DSA also builds more rigour and accountability into obligations around risk management. The primary way in which the DSA is more robust than the OSA is in the transparency of reporting and audit mechanisms which counter the potential subjectivity of the risk assessment approach. The European Board for Digital Services, together with the Commission, must publish annual reports identifying the most prominent and recurrent systematic risks reported by platforms together with best practices to address these risks. As a result of the particular focus on the largest providers, measures designed to ensure accountability can also be more onerous. For example, providers must pay for an independent audit to assess their compliance and provide access to data for both the Commission and also approved researchers. They must put in place a compliance function and 6-monthly transparency reports must be produced in line with requirements set out in Article 42. The DSA also specifies that risk assessments should be carried out annually, and when any new functions are implemented that are likely to have a critical impact on the risks identified. This demonstrates the appeal of a tiered approach, under which greater accountability and more robust oversight can be expected of those companies with greater reach, risk and resource.

The OSA's requirement for companies to complete a risk assessment focused specifically on children is not matched by corresponding accountability measures which might better incentivise redesign of services. A tiered approach, as seen in the DSA, would enable this by imposing the greatest regulatory burdens on the companies presenting the greatest risk, through measures such as require publication of risk assessments, tougher transparency reporting or independent researcher access to data (which in turn could in turn inform Codes of Practice). These changes would shift the focus firmly towards visibility and oversight of companies' systems rather than content and might also impact their public reputation, all likely to incentivise better outcomes (Table 2).

Reliance on age checks

In addition to these more general problems of proportionality and accountability, one other aspect of the OSA seems particularly challenging to enact. In order for companies to fulfil their duty of care towards children, they need to know which of their users are children, and this means assessing age. Section 12(4) states that “the duty set out in subsection (3)(a) requires a provider to use age verification or age estimation (or both) to prevent children of any age from encountering primary priority content that is harmful to children which the provider identifies on the service”. Section 12(6) goes on to state that the age verification or estimation must be “highly effective” in determining whether the user is a child. This is a critical part of the Act but only limited definitions are provided (e.g., “Age verification” means any measure designed to verify the exact age of users of a regulated service” (OSA 2023, s.230(2)) and Ofcom will provide further guidance in their Codes of Practice. This leaves a number of questions open.

TABLE 2 Table comparing use of risk assessments under the OSA and the DSA.

	OSA	DSA
Applicable to	Providers of regulated services which include user-to-user and search services and pornographic services.	Intermediary services provided to service recipients (whether businesses or consumers) established or resident in an EU member state, irrespective of where the service provider is established. Intermediary services include mere conduit services, caching services and hosting services.
Focus on children	User-to-user service providers must conduct a risk assessment about illegal content and a child access assessment to see if children can access all or part of a service. If so, they must also conduct a children's risk assessment.	Providers of online platforms accessible to minors shall put in place appropriate and proportionate measures to ensure a high level of privacy, safety, and security of minors, on their service.
Review requirements following risk assessment	Within 3 months of publication of guidance + when there is a significant change to any aspect of the design or operation of the service	On date of application and 12 monthly thereafter + before deploying functionalities that are likely to have a critical impact on the risks identified
Age requirements	Risk assessment based on level of risk based on user age.	Risk assessment must take into account the protection of minors and serious negative consequences to the person's physical and mental well-being.
Type of content covered	Illegal content and harmful content	Illegal content—legal but harmful content not defined and no obligation to remove, e.g., self harm, bullying
Systematic considerations	Must analyse how the design and operation of the service (including the business model, governance, use of proactive technology, measures to promote users' media literacy and safe use of the service, and other systems and processes) may reduce or increase the risks identified	VLOPs and VLOSEs must produce an annual assessment of the systemic risks resulting from the design, functioning and use of their services regarding the dissemination of illegal content and use proportionate measures to mitigate identified risks.
Other matters to be considered	Level of risk based on other characteristicsThe different ways in which the service is used, and the impact of such use on the level of risk of harm that might be suffered by children	Dissemination of illegal content, actual or foreseeable negative effects for the exercise of fundamental rights, any actual or foreseeable negative effects on civic discourse and electoral processes, and public security and any actual or foreseeable negative effects in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being.
Mitigation of risks	No requirements specified in OSA, left to guidance	Clear requirements set out (Article 35) + guidance may be provided on specific risks.
Independent audit	No	Yes

TABLE 2 (Continued)

	OSA	DSA
Access of data to researchers	No	Yes
Compliance function required to be implemented by provider	No	Yes
Transparency reports	Yes for category 1/2 services, content to be specified in the notice requesting the report.	Yes, including a report setting out the results of the risk assessment, the audit report and information about the consultations conducted by the provider in support of the risk assessments and design of the risk mitigation measures.

Abbreviations: DSA, Digital Services Act; OSA, UK's Online Safety Act 2023.

First, it matters when assessment of age is to be undertaken. As drafted, service providers can use age verification or estimation, or neither of these if content is appropriate for users of any age. There are key differences between the two, albeit not clearly spelled out in the Bill. Age verification usually requires a higher level of proof ('verification') that a particular individual is a specific age or in a particular age range, whilst age estimation involves estimating whether a user is in a particular age category, often using AI or machine learning (Age Verification Providers' Association, 2023). If age verification is used without additional age assurance checks during service use there is a risk that this is not a robust mechanism. Examples of loopholes spoken about in the Bill's Committee debates included the scenario of a child using their parent's credit card to prove they are over 18 when signing up to a service, but then continuing to use that service without needing to prove this was their card. This is not a purely hypothetical scenario: Ofcom's research on this issue found that a third of those aged 8–17 with a social media profile on at least one platform, had a claimed user age of at least 18 (Ofcom, 2022b).

Second, to understand how feasible and effective the legislation would be, it matters what type of verification and assurance tools are to be used. Service providers are not given clear guidance on this in the legislation, undoubtedly because technology moves too fast and any specific requirements would be quickly out of date. However, given that Ofcom already have a duty to ensure that their Codes of Practice address controls over access to services, having a regularly reviewed list of approved age verification/estimation providers or processes would undoubtedly be useful in ensuring minimum standards of accuracy, privacy and security as recommended by children's charities (5Rights Foundation, NSPCC, CHIS, 2022). This is vital if age checks intended to reduce harms are not themselves to cause harm.

Third, questions arise as to why age verification or estimation is to be employed. The Act requires that in-scope companies use such processes only to identify either that a user is a child or the child's age group. No detail is given as to what range the latter should cover. This may result in significant differences in the level of protection (or restriction) provided to different children as a result of risk assessments unless further defined in the Codes of Practice and subject to regular review. A better approach would involve defining specific age ranges with differing expectations regarding levels of protection, as seen in the UK's Age-Appropriate Design Code, which sets out five different developmental age brackets (Information Commissioner's Office, 2020).

Overall, the concern is that, to ensure the Act's effectiveness, a great deal hinges on the availability of effective, inclusive and privacy-protecting means of checking service users'

age in order to apply additional protections to those identified as children. Much more detail is therefore needed regarding expected practices of age verification and age estimation. This will be provided in subsequent Codes of Practice, but at least the landscape of available options is broader, and potentially more privacy-respecting, than it was when the UK Digital Economy Act first proposed age verification to limit access to online pornography in 2017. Age verification can now be provided by third party organisations, meaning personal data need not be shared with the user-to-user service, whilst token-based systems mean that verification need not be repeated for every new site or service. Age estimation techniques based on user photographs or voice can protect users by deleting data once the check has been completed, albeit still raise concerns about accuracy and fairness. These new techniques may be more costly than the basic self-certification techniques traditionally used by social media platforms (Cooney, 2019) but should be more robust. It remains to be seen whether the resulting checks (necessary for all users except on services unlikely to be accessed by children or suitable for all users) will prove effective and nonintrusive.

CONCLUSION

Regulation of digital services and markets is evolving. Legacy frameworks still shape the fundamental approach taken, as is evident from the OSA's inclusion of duties relating to content as well as systems, despite the fact that a truly systems-based approach would cover these within the full range of potential risks and harms. It is perhaps simply that governments, media and regulators are comfortable with the all-too familiar imperatives of content regulation.

Our analysis of the OSA suggests we are seeing a gradual transition, shifting focus from content regulation and liability regimes to a more systemic approach. Despite its flaws, the UK OSA undoubtedly represents an important step forward, demanding that companies take responsibility for design decisions that place users at risk. Ultimately, a wider focus on how online platforms work rather than a narrow fascination with content is more likely to enable regulators to adapt and respond to new risks and will ensure that responsibilities lie with those able to effect change. However, for this to be effective there must be more clarity in relation to the actual requirements, from age checks through to systems design. Seen in this light, the Act's practical weaknesses can be read as a failure to fully embrace the challenges of regulating systems rather than content, or at the very least, a failure to fully codify this approach, leaving much of the detail to guidelines issued by the regulator. While this approach undoubtedly provides increased flexibility in the face of rapid technological innovation, it places a great deal of control in the hands of a regulator, albeit one that is experienced and well-trusted. Whether this, in practice, is a bug or a feature of the law remains to be seen.

In this article we identified three particular weaknesses of the OSA which may limit its efficacy: two which relate directly to the specific challenge of regulating systems, and one which is a feature of the focus on child safety. In relation to the difficulties of designing systems-based regulation that is both proportional and effective, it will be interesting to observe the natural experiment arising in the counter-position of the DSA and the OSA. Although the OSA defines proportionality, the absence of any equivalent to the DSA's tiered approach reflects an under-estimation of the ways in which the business models of the largest data-rich platforms can exacerbate risk. Relatedly, the OSA lacks the teeth of the DSA when it comes to holding such platforms fully to account. In practice, the largest service providers in the UK will already be required to comply with a range of legislative and self-regulatory requirements across jurisdictions. A number of regulatory lessons are likely to emerge from the natural experiment of these different frameworks, but it will also be

important to acknowledge overlapping requirements and identify areas where information could be shared, such as the independent audits produced under the DSA.

In relation to the challenges of age verification and estimation, we await Codes of Practice which will provide guidance to companies about implementing this responsibility. The OSA tries to maintain a balance, introducing age verification to underpin risk assessments whilst avoiding presenting them as obligatory age gates to prevent children accessing services. It is unclear whether or how this compromise will work in practice, and concerns about lack of clarity on age verification have been highlighted throughout the progress of the Act (UKIE, 2022). Notably, there has been little public debate about the merits of moving towards a scenario where all of us have to undergo multiple checks to determine whether we are an adult or a child when using apps and services online, with all the opportunities for additional data collection that might entail. Most worryingly, the risk of over-exclusion remains; children may simply be excluded from services to avoid additional costly obligations, whilst adults unprepared, or unable, to verify their age may also find themselves shut out. Whilst there is much to be admired in the goal of an age-appropriate Internet, achieving the right balance between protection and participation is perhaps the most challenging aspect of the Act as it stands.

The sheer length and political turmoil of the OSA's legislative process has itself been problematic, demonstrating only too vividly how rapidly evolving technological contexts challenge old models of good democratic governance. Similarly, the emergence of new risks, such as those associated with the public roll-out of generative AI tools, emphasises the need for more adaptable, future-proof legislation and the desirability of developing more agile governance processes. Ultimately, the UK OSA has tried to embrace an ambitious new model of regulation that could rebalance corporate conduct amongst the biggest technology platforms in favour of the public interest. This is a laudable goal. Protection of that public interest now rests with the UK regulator, whose experience and significantly expanded level of resource may yet, we hope, make all the difference.

ACKNOWLEDGMENTS

No funding was received for this research.

ORCID

Victoria Nash  <http://orcid.org/0009-0001-9161-0730>

ENDNOTES

¹ Consultation at a glance: Our proposals and who they apply to (<https://www.ofcom.org.uk/>).

² Annex-5-illegal-harms-consultation.pdf (<https://www.ofcom.org.uk/>).

REFERENCES

- 5Rights Foundation. (2021). Ambitions for the Online Safety Bill. https://5rightsfoundation.com/uploads/Ambitions_for_the_Online_Safety_Bill.pdf
- 5Rights Foundation, NSPCC, CHIS. (2022). Children's Charities Amendments for the Online Safety Bill (OSA32). <https://bills.parliament.uk/publications/46593/documents/1840>
- Age Verification Providers Association. (2023). Defining assurance, verification and estimation. <https://avpassociation.com/definitions/>
- Angel, M. P., & Boyd, D. (2024). *Techno-legal solutionism: Regulating children's online safety in the United States*. In CSLAW'24: 3rd ACM Computer Science and Law Symposium, March 12–13, 2024, Boston, USA.
- Balkin, J. M. (2020). The fiduciary model of privacy. *Harvard Law Review*, 134(9), 11.
- Buiten, M. C. (2021). The digital services act from intermediary liability to platform regulation. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, 12(5), 361–380.
- Cavoukian, A. (2009). Privacy by design: The 7 foundational principles. Information and Privacy Commissioner of Ontario, Canada (pp. 5, 12).

- Cooney, A. (2019, May 23). The digital age of consent, one year on. *LSE Media Blog*. <https://blogs.lse.ac.uk/medialse/2019/05/23/the-digital-age-of-consent-one-year-on/>
- Demos. (2022). Evidence on the Online Safety Bill (supplementary submission) (OSA92). <https://bills.parliament.uk/publications/47083/documents/2049>
- eSafety Commissioner (Australian Government). (2019). Safety by design – principles: Placing user safety at the forefront of online service design. <https://www.esafety.gov.au/sites/default/files/2019-10/SBD%20-%20%20Principles.pdf>
- Flew, T. (2021). *Regulating platforms*. Polity Press.
- Flew, T., Martin, F., & Suzor, N. (2019). Internet regulation as media policy: Rethinking the question of digital communication platform governance. *Journal of Digital Media & Policy*, 10(1), 33–50.
- Fukuyama, F., & Grotto, A. (2020). Comparative media regulation in the United States and Europe, *Social media and democracy: The state of the field, prospects for reform* (pp. 199–219). Cambridge University Press.
- Gillespie, T. (2015). Platforms intervene. *Social Media + Society*, 1(1), 1. <https://doi.org/10.1177/2056305115580479>
- Gillespie, T. (2018). Regulation of and by platforms. In J. Burgess, A. Marwick, & T. Poell (Eds.), *The SAGE handbook of social media* (pp. 254–278). SAGE.
- Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, 22(6), 854–871. <https://doi.org/10.1080/1369118X.2019.1573914>
- Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(9), 1. <https://doi.org/10.1177/2053951719897945>
- De Gregorio, G. (2022). *Digital constitutionalism in Europe: Reframing rights and powers in the algorithmic society*. Cambridge University Press.
- Helberger, N. (2020). The political power of platforms: How current attempts to regulate misinformation amplify opinion power. *Digital Journalism*, 8(6), 842–854.
- Information Commissioner's Office. (2020). Age appropriate design: A code of practice for online services. <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/age-appropriate-design-a-code-of-practice-for-online-services/#:~:text=The%20code%20is%20a%20set,designing%20and%20developing%20online%20services>
- Kosseff, J. (2019). *The twenty-six words that created the Internet*. Cornell University Press.
- Livingstone, S., & Stoilova, M. (2021). The 4Cs: Classifying online risk to children. (CO:RE Short Report Series on Key Topics). https://www.ssoar.info/ssoar/bitstream/handle/document/71817/ssoar-2021-livingstone_et_al-The_4Cs_Classifying_Online_Risk.pdf
- Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11), 4366–4383.
- Napoli, P., & Caplan, R. (2017). Why media companies insist they're not media companies, why they're wrong, and why it matters. *First Monday*, 22, 5.
- Nash, V. (2019). Revise and resubmit: reviewing the 2019 online harms [White Paper]. *Journal of Media Law*, 11(1), 18–27. <https://doi.org/10.1080/17577632.2019.1666475>
- Ofcom. (2021). Video-sharing platform guidance: Guidance for providers on measures to protect users from harmful material. Accessed July 11, 2024. <https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-1-10-weeks/216479-vsp-harm-guidance/associated-documents/vsp-harms-guidance.pdf?v=326973>
- Ofcom. (2022a, October 22). Ofcom's first year of video-sharing platform regulation. https://www.ofcom.org.uk/data/assets/pdf_file/0032/245579/2022-vsp-report.pdf
- OfCom. (2022b, July 14). Children's online user ages quantitative research study. https://www.ofcom.org.uk/data/assets/pdf_file/0015/245004/children-user-ages-chart-pack.pdf
- Perrin, W., & Woods, L. (2019). Online harm reduction – A statutory duty of care and regulator. Carnegie Foundation. <https://www.carnegieuktrust.org.uk/project/harm-reduction-in-social-media/>
- Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*, 20(1), 293–310. <https://doi.org/10.1177/1461444816661553>
- Price, L. (2021). Platform responsibility for online harms: Towards a duty of care for online hazards. *Journal of Media Law*, 13, 238–261. <https://doi.org/10.1080/17577632.2021.2022331>
- Rojaszczak, M. (2022). Online content filtering in EU law – A coherent framework or jigsaw puzzle? *Computer Law and Security Review*, 47, 105739. <https://doi.org/10.1016/j.clsr.2022.105739>
- Slavtcheva-Petkova, V., Nash, V. J., & Bulger, M. (2015). Evidence on the extent of harms experienced by children as a result of online risks: Implications for policy and research. *Information, Communication & Society*, 18(1), 48–62.
- Suzor, N. (2018). Digital constitutionalism: Using the rule of law to evaluate the legitimacy of governance by platforms. *Social Media + Society*, 4(3). <https://doi.org/10.1177/2056305118787812>

- Theil, S. (2019). The online harms [White Paper]: Comparing the UK and German approaches to regulation. *Journal of Media Law*, 11(1), 41–51. <https://doi.org/10.1080/17577632.2019.1666476>
- UKIE. (2022, September 21). Written evidence submitted by the UK Interactive Entertainment Association (OSA18).
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., & Kenton, Z. (2021). Ethical and social risks of harm from language models. *arXiv*, 2112, 04359.
- Williams, J. (2018). *Stand out of our light: Freedom and resistance in the attention economy*. Cambridge University Press.
- Woods, L. (2019). The duty of care in the online harms [White Paper]. *Journal of Media Law*, 11(1), 6–17. <https://doi.org/10.1080/17577632.2019.1668605>
- Zuboff, S. (2019). *The age of surveillance capitalism*. Profile Books.

How to cite this article: Nash, V., & Felton, L. (2024). Treating the symptoms or the disease? Analysing the UK Online Safety Act's approach to digital regulations. *Policy & Internet*, 1–15. <https://doi.org/10.1002/poi3.404>