GOV.UK

Department for
Science, Innovation
& Technology

Research and analysis

Potential impact of the Online Safety Bill

Published 23 October 2024

Contents

This publication is available at https://www.gov.uk/government/publications/potential-impact-of-the-online-safety-bill/potential-impact-of-the-online-safety-bill

Method and background

Research objectives

Revealing Reality carried out a piece of work in 2019 to estimate the number of organisations in-scope of the Online Safety Bill (then Online Harms White Paper (OHWP)) and explore the likely costs of compliance for those in-scope. This included a range of organisations – covering different sizes of organisation, a mixture of types of organisation and organisations with different levels of risk. The findings from this research informed the regulatory impact assessment which was published when the Bill was introduced to Parliament.[footnote 1]

However, since then there have been some changes to the policy and the requirements on in-scope platforms. We have been commissioned by the Department for Science, Innovation and Technology (DSIT), the government department responsible for the Online Safety Bill,

to gain an updated understanding of the likely impact. This will feed into DSIT's updated impact assessment which will be published after the legislation is enacted.

Policy changes since the 2019 work

Since the OHWP, and as a result of extensive engagement, there have been numerous changes to the policy. Changes made to the Bill following pre-legislative scrutiny include:

- The list of priority offences, a list of the most harmful offences which companies must take proactive measures to prevent individuals from encountering via their service, in primary legislation.
- Additional duties on Category 1 services (the largest in-scope user-to-user services) to ensure adult users are given the option to verify their identity, and empowerment tools to have more control over the legal content that they see as well as who they interact with.
- A standalone provision that requires all service providers that publish or display pornographic content on their services to prevent children from accessing this content.
- A new duty requiring Category 1 services, as well as large search services (Category 2A), to put in place proportionate systems and processes to prevent users encountering fraudulent adverts on their service.
- New duties requiring Category 1 services to protect news publishers' content, journalistic content and content of democratic importance.

- Duties on Category 1 services to assess their impact on free expression and privacy when adopting safety policies.

Changes made to the Bill following its introduction to Parliament include:

- Category 1 organisations will no longer be required to assess risks from and set terms of service in relation to legal but harmful content and activity accessed by adults.
- Category 1 organisations will be required to set clear terms of service in relation to the restriction or removal of user-generated content, and the suspension or banning of users on grounds related to user-generated content. These terms of service must be clear, easy to understand and consistently enforced.

The interviews with organisations for this work were conducted between March and June 2023, and were based on the Bill as introduced in the House of Lords in January 2023. There have been a number of changes to the Bill following its introduction to the Lords which are not reflected in this research. For example:

- Publishers of pornographic content will need to use age verification or age estimation to ensure that children are not normally able to encounter pornography on their service. User-to-user providers which allow pornography under their terms of service will also need to use age verification or age estimation to prevent children encountering pornography identified on their service.[footnote 2]

This project aims to:

- Understand the current approaches taken by in-scope organisations to identify and prevent harm.
- Quantify the resources and costs behind their approach to mitigating online harm.
- Explore how this may change if a duty of care were enforced – using the updated requirements as outlined in the Online Safety Bill

Sample and recruitment

This work set out to recruit a strategic sample of in scope organisations providing a variety of online services, which will be affected by the regulation in different ways.

Primary sampling criteria

The primary criteria for selecting organisations to take part were:

Risk level. The work set out to include a range of platforms likely to fall within the Online Safety Bill's (OSB) Category 1, 2A and 2B framework.

> Categorisation of in-scope organisations as outlined in the Online Safety Bill[footnote 3]
>
> Category 1: the largest user-to-user services with the greatest influence on public discourse due to their functionalities and number of users, and other relevant characteristics and factors. They will have overarching transparency, accountability and free speech duties, as well as a range of other duties, including in relation to transparency, user empowerment, news publishers and journalistic content, content of democratic importance and fraudulent advertising,
>
> Category 2A services: the largest search services, with transparency and fraudulent advertising duties.
>
> Category 2B services: user-to-user services with a specified number of users, functionalities and other relevant characteristics and factors, with transparency requirements, but no other additional duties.

Researchers used the 'tier system' developed in the previous work to classify organisations based on risk. The tier system accounts for the number and type of in-scope features a platform has and it can be used as a proxy for risk. The tier system does not map perfectly onto the categories within the OSB, as it predates these categories. The tiers provide an indication of level of risk posed by a platform, but not all 'high risk' platforms under the tier system will necessarily be designated as Category 1, 2A or 2B.

The research set out to recruit:

- 15 x Tier 2 organisations (lower risk – e.g. forum, gaming platform)
- 15 x Tier 3 organisations (highest risk – e.g. most social media platforms)

Tier 1 organisations (lowest risk – e.g. a retail site with a reviews function) were not included in this research. Although there may be a number of Tier 1 organisations in scope of the regulations, given the regulation will be proportionate to the risk a platform poses, the requirements on them will be much less. Furthermore, many of the changes / additional requirements that have been introduced will be placed on larger / higher risk platforms, so we have chosen to focus on these in our sample.

Size: To include organisations of a range of sizes: (micro, small, medium, large) where possible, given the impact is likely to differ based on this. The following definitions were used to classify an organisation's size:

- Micro: 0-9 employee
- Small: 10-49 employees
- Medium: 50-249 employees
- Large: 250+ employees

Secondary sampling criteria

There are additional factors that may affect an organisation's approach to tackling online

harm, and their resources to do this. Across the sample, aim to include:

- A range of types of online platforms – including organisations from a range of sectors such as social media, gaming, search, dating apps and peer to peer marketplaces.
- Organisations aimed at audiences / likely to have users who are more likely to be vulnerable e.g. children
- Type of business – to include different types of business including charities and limited companies
- Revenue – the organisations' revenue

Recruitment

Recruitment efforts targeted 25 Tier 3 organisations and 44 Tier 2 organisations, all of which were contacted directly via the most relevant member of staff in relation to trust and safety, compliance and public policy.

To ensure a diverse sample, organisations across 27 different industry sectors were approached:

- Social media
- Video Sharing
- Instant messaging
- Professional networking
- Gaming
- Forums
- Q&A
- Pornography
- Video calling
- Search engine
- P2P marketplace

- Marketplace
- Telecommunications
- Photo sharing
- Live streaming
- Dating
- Reviews
- Accommodation searching
- Fitness
- Investment
- Crowdfunding
- Rentals
- Transport
- Sports
- Cloud content management
- Literature
- Genealogy

Final sample of organisations

10 organisations were interviewed as part of this research.

Table 1: Number of organisations interviewed

Tier 2	Tier 3	small (size)	medium (size)	large (size)
6	4	1	4	5

These 10 organisations cover the following sectors:

- Pornography
- Social media
- Reviews
- Video sharing platform
- Search
- Crowdfunding
- Gaming
- Forum
- Job / skills marketplace

Most organisations only provided one service, so interviews were conducted based on that one service provided. However, a few organisations provided

multiple services or platforms, so here interviews covered multiple platforms.

Method

Revealing Reality contacted organisations using a mixture of Revealing Reality's existing relationships from previous work, DSIT's relationships, and contacting individuals on LinkedIn and email.

Revealing Reality then carried out 45–60-minute remote interviews with organisations from March to June 2023. These interviews have been carried out with individuals in a range of roles depending on the size of the business. Roles include Head of Policy and Public Affairs, Head of Operations, Head of Customer Experience, CEO, and Head of Compliance.

Interviews followed a semi structured discussion guide (see the full guide in Annex 1) and covered:

- General background to the organisation and its in-scope activities.
- Processes to prevent, identify and mitigate online harms.
- Costs & resources associated with preventing harms.
- How these costs and resources would likely change under requirements outlined by the OSB. This involved going through the most up to date requirements on in-scope organisations, as outlined by the OSB, to gauge:
  - The extent to which organisations already have processes in place to meet that requirement.
  - For those that don't, how easy/difficult this would be to implement.
  - Costs associated with complying with that requirement.

Each organisation was provided with a detailed information sheet. This is included as Annex 2. The information covered the following things:

- Explanation of the research
- Explanation of the organisation's involvement
  - Length and format of interview
  - Overview of topics covered (see above)
- Detail on how they could prepare for the interview
  - Gathering information and documentation relating to the topics being explored
  - Gathering any figures or stats on the costs and resources put towards protecting users from harm
- Detail on who Revealing Reality are and contact details

- Detail on how their data was to be used
  - E.g. information shared with Revealing Reality will be completely confidential and anonymous (i.e. not connected to any identifiable information about you or your organisation).
- Detail laying out the different duties and responsibilities for in-scope services under the Online Safety Bill as introduced in the House of Lords in January 2023
- Detail of the Harms in scope of the Bill

Limitations of the work

The main limitation to this work is the limited sample. Only three of the largest Tier 3 organisations responded to our requests to take part in the research.

These organisations have a greater number of requirements to comply with, compared to lower risk, Tier 2 organisations, meaning there are certain requirements that only apply to Tier 3 organisations that we have less data on.

Furthermore, some of these organisations are more likely to be affected by requirements such as implementing age assurance technology, as some of these platforms are known to be used by children.

The lack of response from the largest, highest risk organisations could be due to many factors. For example, it may suggest that organisations anticipate limited changes to the cost they are likely to face under the OSB since the previous work was conducted, and therefore there is little reason for them to engage with the work. In the previous work, 11 of the largest, high-risk organisations were interviewed. Indeed, it is likely that the largest companies are already investing in online safety to prepare for the bill.

The findings in this report are based on a limited sample of in-scope organisations that chose to respond to the research request. Therefore, insight and conclusions from this work will present a limited view of the cost implications for in-scope businesses. However, the insight from organisations that did take part, appears to support our hypothesis that organisations anticipate limited changes to the cost they are likely to face under the OSB. Whilst some organisations were able to provide a more comprehensive breakdown of additional costs to them when complying with the duties in the Bill, many were unable to give specific numerical estimates of money or time – and instead reflected more broadly about implications about the upcoming bill. There are likely multiple reasons for organisations' inability to provide numerical estimates, including ambiguity about what the requirements would involve.

There is also a possibility that at this stage, businesses might not understand the current regulation fully, and as a result the costs within this research might not be accurate.

Finally, it's important to repeat that this research captures the views of in-scope organisations based on the Bill as introduced in the House of Lords in January 2023. There have since been further changes to the Bill which may not be reflected in the views presented in this research.

Overarching findings

Generally, organisations did not raise significant concerns, though they need to see the detail of the actual regulatory requirements to fully assess the impact on costs. Most either felt confident that they already operated in a way that would meet most of the requirements on their organisation, or that additional work and changes would be relatively minimal, feasible, and not present a cost that would pose significant problems to their organisation.

Most organisations already had systems in place that they believed would meet the majority of requirements outlined in the OSB. There were multiple factors driving organisations to invest in online safety including:

- Voluntary changes due to shifts in attitudes towards online safety such as:
  - a desire to create a positive environment for users, to retain existing users and attract new users
  - to remain competitive in the industry and keep up with competitors
- pressure from the public, employees and media to improve online safety
- Regulation outside the UK, such as having to comply with the European Digital Services Act and other European regulation that was already in place. Indeed, one organisation mentioned already incurring more cost and administrative burden to adhere to the DSA, than anticipated as a result of the Online Safety Bill.
- Regulation in the UK. Many have been preparing for the Online Safety Bill over the past few years, and made adjustments to account for the bill.

Several also commented that while some compliance actions may involve new costs and resource allocation, they are happy to continue investing in safety. The exception to this is where organisations saw compliance with requirements as purely administrative or bureaucratic, in which case resource spend was seen as a negative burden.

While these administrative costs were generally considered likely to be low, the ambiguity around certain requirements raised some larger concerns about cost and viability. These were mostly relating to ambiguity in how

legislation might be interpreted and applied by the regulator. This ambiguity contributed to platforms being unable to provide cost estimates, as they were unsure what regulation would look like in practice.

Detailed information relating to costs on requirements outlined in the OSB

It should be noted that these findings are based on a relatively small number of interviews conducted. Findings are caveated throughout where they are based on particularly small numbers of platforms responses.

Whilst some organisations were able to provide a more comprehensive breakdown of additional costs to them when complying with the duties in the Bill, many were unable to give specific numerical estimates of money or time. We have included all insight relating to additional costs in the table below.

For each requirement of the Online Safety Bill, we have indicated where organisations anticipated:

- No additional costs were anticipated, and any reasons given for this
- Staffing / resource costs
- Service-specific costs (e.g. buying a service from an external organisation)
- Investment in technology / infrastructure of the platform
- Reduction in users / revenue
- Opportunity costs

It’s important to note that most of these cost estimates given below are not cross referenced. They are what the platforms reported based on their own research and estimates. They are by no means representative figures.

Requirement	Estimated costs and reflections
Risks and assessments A risk assessment of illegal content that may appear on their service, ranging from online fraud to terrorism. Services that are likely to be accessed by children will also have to	**Minimal additional cost expected since previous assessment:** Tier 3 organisations had risk assessments in place already. Most were confident they could tweak elements of these to comply with Ofcom’s requirements, under a reasonable interpretation of the bill, therefore representing a minimal additional cost.

Requirement	Estimated costs and reflections
do a risk assessment concerning content which is harmful to children.	Some of the Tier 2 organisations did not currently have risk assessments in place, but were generally willing to do so and felt this would not be too onerous. Indeed, one micro Tier 2 organisation had already planned to start doing risk assessments. These organisations assumed that their existing compliance or online safety team would absorb the responsibility to complete risk assessments into their role. **Staffing / resource costs:** One Tier 2 micro organisation estimated it would take approximately 1 week to complete and cost £1,500 in staff time. This cost estimation does not include maintaining the risk assessment and keeping it up to date. **Opportunity costs:** This Tier 2 organisation highlighted the opportunity cost associated with this, as a 'start-up', as it meant the individual completing the risk assessment would not be able to pursue new business for the platform for a week.
Illegal content	**No additional cost expected:** All organisations assumed they were already doing what they needed to do to minimise the risk of, identify, and respond to illegal content on their platforms, and were not anticipating additional costs. **Staffing / resourcing costs:** One Tier 2 micro organisation operating a search platform highlighted that their work was currently focused on identifying and preventing one type of illegal content, CSEA, as they believed there was a limited risk of other types of illegal content appearing on their platform. They were aware they may

Requirement	Estimated costs and reflections
	have to expand their capability (their current costs to do a risk assessment are set out above at approx. £1500 in staff time) to address additional illegal content in risk assessments. Though they did not give specific cost estimates for this.
Child safety	**No additional cost expected:** Most platforms had policies about the minimum age to use the platform and varying degrees of age assurance in place, that they thought were sufficient. **Service-specific costs:** One large gaming organisation received indicative costs from a third-party age assurance provider which would be 10p per user verification, plus 2-3 months of 4-5 developer's time to implement the technology. They receive 16, 000 sign-ups a day so 10p per verification would represent a significant cost if they were required to implement it for new users as well as all currently registered accounts (350 million accounts).[footnote 4] Note: It's important to note that these cost estimates are not cross referenced. They are what the platforms reported based on their own research and estimates. They are by no means representative figures **Reduction in users / revenue:** A few organisations spoke about the issue of additional friction that would be introduced if a platform implemented age assurance technology i.e., additional steps or clicks to access the platform, which may put users off using their site and represent a reduction in revenue as a result of a reduction of users, or drive existing users to alternative platforms that are not compliant with age assurance regulation.[footnote 5]

Requirement	Estimated costs and reflections
Term of service	**No additional cost expected:** All- organisations interviewed already have published terms of service. No organisation anticipated having to make major changes to their terms of service. And all organisations with services likely to be Cat 1 reported foreseeing no issues with complying with the accompanying duties that comes alongside enforcing their terms of service. All organisations felt confident they were already able to enforce their terms of service, and wouldn't require additional moderation to do so **Investment in technology** When the adult safety duties (adult 'legal but harmful') were part of the Bill, one large gaming organisation explored the cost for AI based chat moderation which was going to cost £450,000 to buy from an external third party, plus three to four months for their five-person developer team to integrate it. This was perceived as a large cost burden, but one they are no longer considering given this measure as the adult safety duties have been removed. **Investment in technology:** One large sized gaming organisation relies on post to receive complaints from individuals, as they get a lot of 'trolling' complaints, in their words, that they want to reduce through adding friction to the reporting process. If they were required to upgrade this to a web form, this might reduce friction and they were concerned it could lead to more reports. This would require a small investment of time from developers (no direct estimates given). But would likely increase staff time to handle these complaints as well. As they can't predict the increase in complaints, they couldn't put

Requirement	Estimated costs and reflections
	direct cost estimates to this increased staff time.
Reporting Child Sexual Exploitation and Abuse (CSEA)	**No additional cost expected:** Most organisations were already reporting Child Sexual Exploitation and Abuse (CSEA), either to the National Center for Missing & Exploited Children (NCMEC) or the National Crime Agency (NCA), or Internet Watch Foundation (IWF). Where organisations were headquartered in the US, they were reporting to NCMEC due to pre-existing requirements to do so, while others said they were or would voluntarily report to the NCA / IWF. One medium sized crowdfunding organisation did not currently have specific reporting processes in place for CSEA, but report anything illegal or criminal to the NCA, so would expect there to be little or no additional resource required there. **Service-specific costs:** Some paid to the IWF for access to their URL list One large pornography organisation reported using AI / hash-list scanning to detect underage material, which they reported cost 5-20p per check i.e. to check a piece of material. However, it is unclear where this cost has arisen from.
Transparency reports	**No additional cost expected:** Most Tier 3 organisations were already producing transparency reports, and under a ‘reasonable interpretation’ of the bill, did not believe that the requirement would add significant cost to the organisation. However, they wanted greater detail and clarity about what transparency reports were likely to include in order to make an

Requirement	Estimated costs and reflections
	assessment about how challenging this would be to produce. **Service-specific costs:** Tier 2 organisations were less likely to have transparency reports in place already, although anticipated this would be feasible should it just require the reporting of data that the organisations already had. One medium sized crowdfunding organisation raised the challenge that they did not currently have all the data required to write a transparency report feeding into one place, so would need to set up a new system for collecting and collating this data. A few Tier 2 organisations flagged that writing the reports may require effort for 'little reward' given their low number of reports. One micro Tier 2 search organisation estimated it would take about half a day for one employee to complete a transparency report. They did not provide a direct cost estimate for this. **Investment in technology:** Another Tier 2 organisation reported that there are several data points they can't collect from a tech capability point of view. Whilst they are already in discussions with the tech team about how to increase the amount of data available for internal transparency reports, if it was required of them to provide much more than they're currently capable of doing, it could incur significant costs in terms of investment in new technology.
Fraudulent advertising	**No additional cost expected:** Organisations that this requirement was likely to apply to generally felt that they already had due diligence processes in place for any paid-for advertising, and therefore would not face additional costs. However, one organisation

Requirement	Estimated costs and reflections
	raised a challenge that if this related to a user phishing on the platform it would become more complicated.
User verification and empowerment	**No additional cost expected:** Some platforms already had user verification processes in place, for example for content creators on an adult entertainment platform. Note: this duty only relates to Category 1 organisations, and there were few organisations interviewed who are likely to fall into this Category.
Freedom of expression and privacy	**No additional cost expected:** Organisations who expected this duty to be applicable to their platform tended not to raise major concerns, as they felt that their community guidelines already considered freedom of expression. None mentioned carrying out specific impact assessments on freedom of expression, though this was not specifically probed into. Note: this duty only relates to Category 1 organisations, and there were few organisations interviewed who are likely to fall into this Category.
Protected content	**Staffing / resourcing cost:** Our organisation felt they would need to set up a separate team to decide what content shouldn’t be removed from their platform. Note: this duty only relates to Category 1 organisations, and there were few organisations interviewed who are likely to fall into this Category.
Standalone pornography provision	**Staffing / resourcing cost:** One large pornography provider estimated it would cost around 10-20p per user

Requirement	Estimated costs and reflections
	verification.[footnote 6] Using this cost they estimated this would cost the business about £500,000 per month for the UK alone. They felt this would not be feasible. The interviewee did not give detail about how this initial estimate was calculated.. **Reduction in users / revenue:** This pornography provider also raised the issue of additional friction that would be introduced if a platform implemented age assurance technology i.e. additional steps or clicks to access the platform, which may put users off using their site and represent a loss in revenue as a result of a reduction of users, or drive existing users to alternative platforms that are not compliant with age assurance regulation.[footnote 7] Note: this duty only relates to pornography publishers and only one pornography provider has been engaged in this research.

Conclusions

From the interviews conducted in this research, the emerging sense is that most platforms are not overly worried about compliance cost and impact based on what they expect the implications of the Online Safety Bill to be. There are concerns in areas where the ambiguity in how the regulation is applied could have larger ramifications.

This report is based on ten out of the target of thirty interviews, so concern should be exercised in drawing conclusions from these findings. The low response rate from larger tier 3 companies in particular prevents this research from providing a comprehensive picture of the impact of the Online Safety Bill in terms of costs to UK businesses.

Annex 1: Topic guide for organisation interviews

Project objectives

To explore the potential impact that the Online Safety Bill (OSB) will have on UK organisations. This work should explore the gap between what organisations are currently doing, and what they will be required to do under the OSB, assessing the likely cost of regulation. This should provide updated figures from the initial impact assessment research carried out by Revealing Reality in 2019.

Specific objectives for interviews with organisations are to explore:

- The online activities they enable that carry risk of harm to users (as outlined in the Online Safety Bill)
- Their current practices and processes to mitigate that risk and to identify any harm occurring
- Where available, quantification of the associated resources and costs of practices and processes to identify and prevent harm; and (Note: this is the primary research objective)
- How these costs and resources would change under the OSB
- Comparison of costs and resources across different organisation types

Interview set up

- We are an independent social research agency based in London
- As outlined on the information sheet we have sent you, we have been commissioned by the Department for Science, Innovation and Technology (DSIT), the government department responsible for the Online Safety Bill, to explore the different ways in which organisations identify and try to mitigate against the types of online harms outlined in the Bill, and estimate the likely costs that may be associated with this.
  - This work will inform DSIT impact assessment of the Online Safety Bill, ensuring regulation is proportionate and feasible
  - We are talking to a number of organisations whose platforms enable users to interact with each other, or generate and discover user generated contact, as with these types of features, there is a potential risk of harm. An organisation being selected to take part in the research does not necessarily mean they are in-scope of the regulation, or that they will be designated as category 1, 2a or 2b.
- Anything you tell us will be kept confidential – if you are happy for us to, we will provide DSIT with a list of those who have taken part in the research, but we will not attribute individual pieces of data to specific organisations

- If it is okay with you, we will voice record the interviews, so we don't have to make lots of notes whilst on the call, but these recordings will not be shared with anyone outside of the research team at Revealing Reality or used for any purpose other than writing up notes
- You do not have to answer any questions that you do not want to. We also recognise you may not be able to answer all of the questions, and have suggested that we speak to different people in different teams where appropriate
- If you have any questions at any point do not hesitate to ask
- The call will last around 45 minutes-1 hour

Topics to cover

- General background to the organisation and its in-scope activities
- Processes to prevent, identify and mitigate online harms
- Costs & resources associated with preventing harms
- How these costs and resources would likely change under requirements outlined by the OSB

Note: Revealing Reality will have a significant amount of information about organisation's approaches to online harm, from previous interviews for DSIT (formerly Department for Digital, Culture, Media and Sport (DCMS)) and reviewing their responses to Ofcom's call for evidence. This is particularly the case for Tier 3 organisations.

Interviewers will have this information to hand so that conversations build on prior knowledge (e.g. about existing approaches to tackling online harm, resources associated with this and how they anticipate this changing under a duty of care).

Questions will be tailored to each organisation based on existing evidence that researchers have.

Questions will also be adapted based on whether the organisation is likely to be classified as a Category 1, 2A or 2B under the OSB.

Lines if pushed on categorisation:

At this stage, we have made clear the factors that will be taken into consideration for Category 1 status. The exact thresholds are still to be set by the government, following advice from the regulator, after passage of the legislation. This is to ensure the process is objective and evidence-based. Ofcom will then be required to assess services against these thresholds and publish a register of all those which meet both thresholds. We expect the largest social media platforms to all be Category 1 services.

General understanding of the organisation and prevalence of online harm

Tell me a bit about your organisation (both UK focused and international context) Note: much of this information to be ascertained through desk research. Only ask if not publicly available (e.g. for smaller / tier 2 organisations)

- What services does your organisation provide?
- What types of features and functionality exists that enable UGC (user-generated content) or P2P interactions
  - How core to the organisations' functioning are these features?
- Who are the main users of your service?
  - PROBE: whether services likely to be accessed by children, including age restrictions
- Researcher to sense check / gather data on the organisation's:
  - Size, number of employees
  - Revenue
  - Reach
  - Business model and revenue generation

> We sent a list of the types of content (illegal and harmful) that platforms will be required to address under the Online Safety Bill, which I am sure you are familiar with. To what extent are these types of content / harms something that your platform is already tackling?

- Which harms are most prevalent on your platform?
  - Are there any additional harms that your organisation is more focused on tackling? Which?
  - Are there any categories of harm that your current systems and processes do not proactively seek to identify?
- On which parts of your platform/activities are you most likely to observe harm occurring?

Illegal content that platforms will need to remove includes

- child sexual abuse
- controlling or coercive behaviour
- extreme sexual violence
- fraud
- hate crime
- inciting violence
- illegal immigration and people smuggling
- promoting or facilitating suicide

- promoting self harm
- revenge porn
- selling illegal drugs or weapons
- sexual exploitation
- terrorism

harmful content

Some content is not illegal but could be harmful or age-inappropriate for children. Platforms will need to prevent children from accessing it.

Harmful content that platforms will need to protect children from accessing will include:

- pornographic content
- online abuse, cyberbullying or online harassment
- content that does not meet a criminal level but which promotes or glorifies suicide, self-harm or eating disorders
- Which harms are the hardest to mitigate against? Why?

Mitigations and associated costs

> What are some of the processes your organisation has in place to identify and prevent online harm?

Researcher to PROBE based on the requirements for in-scope organisations listed below (notes taken from DCMS' impact assessment), if not spontaneously raised by platforms.

For each requirement, researcher to prompt:

- What is your view on this requirement?
- To what extent is this something you already do / have in place?
  - How do these processes work in practice?
  - To what degree are processes focused on specific hazards or operating at a general level?
  - What do these processes cover?
  - PROBE type of content, activity or harm, different users, platform features, information included
  - What works well / less well? Why?
  - What evidence / data do you use to assess how well they are working?
  - What is the balance of proactive vs reactive measures, e.g. moderation?

  - What are the challenges in implementing this?
- [If the organisation does not have this in place] How easy/difficult would it be to do?
  - What would the challenges be in implementing this mitigation?

(PROBE: finances, design limitations, resources, lack of clarity on what to do to prevent the harm, lack of identification of the harm)

List of requirements for in-scope organisations (to be tailored based on the Category of platform / whether they are a pornography publisher):

- Terms of service / community guidelines
  - Assumed most organisations will have this already, but they may need to update this in response to future codes of practice
- Conducting risk assessments
  - All companies in scope will have to carry out an illegal content risk assessment and if 'likely to be accessed by children' to also carry out a children's risk assessment
  - It is no longer a requirement for Category 1 organisations to assess the risk of legal but harmful content, but what is your view on what the implications or impact of this would have been, if included as a requirement?

Content moderation (AI and Human)

- To use proportionate (to the recent risk assessment and size and capacity of the org) measures to effectively mitigate the risk of harm to individuals
- Content moderation is not specifically required, but it is likely that additional human and automated content moderation will be required in order to comply with duties, including making improvements to both of those through greater investment.
- User reporting
  - To provide users with mechanisms to report illegal content or activity and content which is harmful to children
  - Could be as simple as a visible email address
- Complaints procedures
  - Easy to use complaints process, accessible by adults and children, to appeal the wrongful takedown of their content and to raise concerns that a company has failed to fulfil its duties under the Bill
- Employing age assurance technology
  - To say what technology they are using, if any, and show they are enforcing their age limits
  - High risk platforms are likely to adopt this in order to comply with child safety duties

    - Pornography providers to prevent children from accessing published pornographic content
- User verification and empowerment duties (Category 1 organisations only)
    - To offer optional user identity verification
    - To provide empowerment tools to users that give them more control over their online experience. This may include ability to block or restrict who contacts you, and the ability to filter out certain content
- Transparency reporting (Category 1, 2A and 2B only)
    - To publish annual reports on platform
    - harm and related actions taken by the platform
- Fraudulent advertising duty (Category 1 and 2A organisations only)
    - To minimise the publication and/or hosting of fraudulent advertising
    - Likely required to conduct CDD (customer due diligence) on advertisers
- Reporting online CSA to body
    - Cost of detecting CSA content, reviewing and preparing reports and reporting CSA to designated body
    - They will currently not report in the UK but many will report to NCMEC in the US [so worth capturing]
- PROBE for any additional mitigations
    - E.g. decisions to remove / limit certain features that could be risky, banning / restricting users, access to databases such as Photo DNA
- How have the mitigations you have in place changed over time?

What does your organisation have in place to protect free speech?

Requirements on Category 1 organisations only

- FoE and privacy IAs (Category 1 organisations only)
    - To assess the impact of their policies and publish the steps they are taking to protect users' rights to freedom of expression and privacy
- Protected content
    - To put systems and processes in place to protect journalistic content and content of democratic importance when taking action against users or content
- Transparency, accountability and freedom of expression duties
    - To ensure that the terms of service provide sufficient detail for users to understand what content is and is not permitted on the platform
    - Having processes to ensure they only remove or restrict access to content, or ban or suspend users, except where allowed by their terms of service, or where they otherwise have a legal obligation to do so

We are interested in learning more about the costs associated with preventing harm, and protecting freedom of expression, to ensure any new regulation minimises costs to organisations.

Note: It is likely that interviewees may need to go away and find this information or pass us onto someone else to answer some of these questions

- Have you ever tried to establish the cost of some of the mitigations you put in place?
  - Roughly how much has been spent on different ways to identify and mitigate against harm? PROBE for specific mitigations mentioned during the interview:
    - Cost in numbers or as a % of total outgoings / revenue
    - Staff in numbers or as a % of total workforce or in time (staff resource)
    - Cost per user or per report
    - Set-up vs ongoing/operating costs
  - What drives this cost? (PROBE type of content, prevalence/amount of content, complexity in determining illegality or breach of terms, information required)
- How do you anticipate the resources allocated to preventing harm changing under a duty of care? [to probe around this for orgs with pre-existing processes in place and those without them]
  - What additional costs are you anticipating? Probe for:
    - New measures that need to be introduced to address requirements, that you don't currently have
    - Strengthening or changing existing measures and processes due to requirements
    - Increases in volume of reports, moderation or other processes as a result of the requirements
  - PROBE potential additional costs for each requirement, by revisiting any unmet requirements for organisation (accounting for the Category of the organisation)
  - To what extent are these changes a direct result of the regulations, or were you planning to bring these in regardless of legislation?

Final reflections on the impact of the OSB

Do you have any final reflections on how the OSB will impact your organisation?

- Are there any other additional costs to your organisation that may occur as a result of the OSB that have not yet been discussed? (PROBE: Familiarisation with regulations, industry fees, potential enforcement action, the possibility of making information available to the public e.g. areas involving publishing statements)
- Do you have any concerns about any of the requirements?
  - Which requirements do you think will be the most challenging to meet Why?
  - How do you think your organisation will address this?
- Are there any unanticipated consequences of the OSB that you can foresee for your business? Both advantages and disadvantages?
  - PROBE: E.g. gaining or losing customers, greater or poorer customer satisfaction, business reputation, advertising or other revenue changes, etc.

Thank you and close.

Notes of information from the OSB referred to in the interviews

Category of organisation

Category 1 services: the largest online platforms with the widest reach including the most popular social media platforms

Category 2a services: the highest reach search services, with transparency and fraudulent advertising requirements.

Category 2b services: other services with potentially risky functionalities or other factors, with transparency requirements, but no other additional duties

Note of requirements on organisations from previous impact assessment and updated bill:

- Reading and understanding regulations (familiarity)
  - Reading and understanding codes of practice, primary and secondary legislation
- Ensuring users are able to report harm and content they consider to be illegal
  - Could be as simple as a visible email address
  - (also complaints procedures?)
- Updating terms of service

    - Assumed most will have this already, but they may need to update this in response to future codes of practice
- Conducting risk assessments
    - Everyone has to carry out an illegal content risk assessment and if ‘likely to be accessed by children’ to also carry out a children’s risk assessment
- Additional content moderation
    - This is not specifically required, but it is likely that additional content moderation will be required in order to comply with duties
    - Platforms required to use proportionate (to the recent risk assessment and size and capacity of the org) measures to effectively mitigate the risk of harm to individuals
- Employing age assurance technology
    - High risk platforms are likely to adopt this in order to comply with child safety duties
    - Platforms will have to say what technology they are using, if any, and show they are enforcing their age limits
    - Pornographic services (non-UGC) will be required to ensure that children cannot access their services
- Transparency reporting (Cat 1, 2A and 2B)
    - Producing annual published reports on platform harm and related actions taken by the platform
- Fraudulent advertising duty (Cat 1 and 2A only)
    - Likely required to conduct CDD (Customer due diligence) on advertisers
- User verification and empowerment duties (Cat 1 only)
    - This relates to requirement on large social media platforms to offer optional user verification and provide user empowerment tools for a list of content categories
    - Empowerment tools may include ability to block or restrict who contacts you, able to filter out certain content
- FoE and privacy IAs (Cat 1 only)
    - Publishing assessment of impacts on FoE and privacy
- Reporting online CSA to body
    - Cost of detecting CSA content, reviewing and preparing reports and reporting CSA to designated body
- Industry fees
    - Ofcom’s operating costs paid by industry fees (tiered and there will be a threshold for which a platform has to pay this fee)
- All SMEs exempt
- Enforcement action (fines and business disruption measures)

- Fines can be issued for failing to comply with their duties – up to £18 million or 10% of qualifying global turnover, whichever is higher
- Business disruption measures – remove third party services like advertising
- Senior managers can be held criminally liable, and face jail, a fine or both, for failing to ensure the company complies with Ofcom's information requests

Notes on content and harms in-scope of the bill:

Illegal content that platforms will need to be tackled

Priority illegal content is content that amounts to an offence on a list of the most harmful offences which companies must take proactive measures to prevent individuals from encountering via their service and to minimise the length of time for which any such content is present. These offences include:

- Terrorism offences
- Child sexual exploitation and abuse offences
- Encouraging or assisting suicide
- Offences relating to sexual images i.e. revenge and extreme pornography
- Incitement to and threats of violence
- Hate crime
- Public order offences - harassment and stalking
- Drug-related offences
- Weapons / firearms offences
- Fraud and financial crime
- Money laundering
- Exploiting prostitutes for gain
- Organised immigration offences and human trafficking
- Coercive or controlling behaviour
- Foreign interference offence

The full list of priority offences can be found in Schedules 5 (terrorism), 6 (child sexual abuse and exploitation) and 7 (priority offences) of the Online Safety Bill. Note: the government has also committed to adding coercive or controlling behaviour, Section 24 of the Immigration Act 1971 and Section 2 of the Modern Slavery Act, and the new Foreign Interference Offence (legislated for in the National Security Bill) to this list.

Companies also have to remove any other illegal content where there is an individual victim, where it is flagged to them by users or they become aware of it through any other means.

Content that is harmful to children

Companies will need to take specific action to prevent children from encountering content that has been designated as ‘primary priority’ harmful content to children and must take an age appropriate approach to protecting children from ‘priority’ harmful content.

Primary priority content (children must be prevented from encountering altogether):

- Pornography
- Content promoting self-harm (with some content which may be designated as priority content, e.g. content focused on recovery from self-harm)
- Content promoting eating disorders (with some content which may be designated as priority content, e.g. content focused on recovery from an eating disorder)
- Legal suicide content (with some content which may be designated as priority content, e.g. content focused on recovery)

Priority content (companies need to ensure content is age appropriate for their child users):

- Online abuse, cyberbullying and harassment
- Harmful health content (including health and vaccine misinformation and disinformation)
- Content depicting or encouraging violence

Content in scope for the user empowerment duties:

Category 1 platforms will be required to provide optional user empowerment tools to give users greater control over the content they see. They will need to provide these tools for:

- Content that encourages, promotes or provides instructions for suicide, self-harm or eating disorders
- Content that is abusive or incites hatred against people on the basis of race, religion, sex, sexual orientation, disability or gender reassignment.

Annex 2: Project information sheet

Thank you for taking part in our research. This sheet provides information about the research. If you have any further questions, please get in touch.

About the project

This research has been commissioned by the Department for Science, Innovation and

Technology (DSIT). The research aims to strengthen DSIT's understanding of the potential economic impact of the proposed Online Safety Bill regulation to organisations.

This work will enable DSIT to update their regulatory impact assessment and ensure the economic impact of the regulation to organisations is proportionate.

Your involvement

Taking part will involve a 45-minute – 1 hour interview with a researcher from Revealing

Reality over Zoom or Teams. We are interested in learning about organisations' current approaches to identifying and preventing online harm, the costs and resources required, and how this may change if a duty of care were enforced.

We are talking to a number of organisations whose platforms enable users to interact with each other, or generate and discover user generated contact, where they may be a potential risk of online harm. An organisation being selected to take part in the research does not necessarily mean they are in-scope of the regulation, or that they will be designated as a Category 1, 2a or 2b organisation.

We will discuss how organisations may respond to requirements on organisations in-scope of the bill, some of which are listed in Annex 1. Note: requirements vary depending on the type and size of the organisation.

How you can prepare

We would really appreciate, ahead of the interview, if you are able to gather any information about what your organisation is currently doing to protect its users against online harm, the costs associated with this, and how you anticipate this may change under a duty of care. Below are some examples of the types of questions that we would like to discuss with you.

- What are some of the harms that your organisations are trying to tackle? Which are the most common on your platform? Which harms are the hardest to mitigate against?
- What are the ways in which you currently mitigate online harm?
- What resources do you allocate to protecting your users from harm and what are the associated costs? This may include hiring staff, investing in technology, producing reports on the prevalence of harms etc.
- How have these costs changed in the last few years?
- How do you anticipate this changing under a duty of care?

Any figures or stats that you are able to prepare on the costs and resources put towards protecting users from harm would be really helpful for us to understand the extent to which organisations are already devoting budget towards preventing harm. We recognise these will be estimates, and we may need to speak to someone in the finance department about this.

Who we are

Revealing Reality is a research company based in London. We specialise in researching complex issues and spending extended periods of time understanding multifaceted industries and organisations. All of our researchers have up-to-date DBS checks and abide by the Market Research Society Code of Conduct.

If you have any further questions about the research, feel free to get in touch. Olivia Nettleton (Associate Director)

- Email: olivia.nettleton@revealingreality.co.uk
- Phone: +44 (0)20 7735 8040
- www.revealingreality.co.uk (https://revealingreality.co.uk/)

If you have any questions for DSIT please contact: soh-analysis-team@dcms.gov.uk.

How your data will be used

Your privacy is extremely important to us.

Any information you share with Revealing Reality will be completely confidential and anonymous (i.e. not connected to any identifiable

information about you or your organisation). Nothing you say will be attributed to you personally or the company you are speaking on behalf of.

The research is entirely voluntary - if at any stage you feel uncomfortable, please do tell the researcher that you would like the session to end.

Revealing Reality will handle your data in accordance with the data protection legislation and we will dispose of any personal information from our system once it is no longer necessary to use.

Annex 3: More information on the Online Safety Bill

Differentiated duties on organisations in scope of the Bill

The Online Safety Bill establishes a differentiated approach to ensure that the duties are proportionate to the risk of harm that different services pose and the capacity of companies.[footnote 8] The table below includes further detail on the duties on services in scope of regulation.

Table 2: Duties on services in scope of regulation.

Duty	Services in scope[footnote 9]
Risk assessment duties: to assess the level of risk on their service from illegal content and activity, and to assess risks for children if the service is likely to be accessed by them.	User-to-user and search services
Illegal content duties: to put in place systems and processes to minimise and remove priority illegal content and to remove non-priority illegal content when identified through user reporting.	User-to-user and search services

Duty	Services in scope[footnote 9]
Child safety duties: if the platform is likely to be accessed by children, to put in place systems and processes to protect children from harmful content.	User-to-user and search services (likely to be accessed by children)
Term of service duties: to have clear and accessible terms of service and not act against users except in accordance with these terms of service.	Category 1 (user-to-user) services - providers with the greatest reach and influence over public discourse
User reporting and complaints procedures:to provide mechanisms to allow users to report harmful content or activity and to appeal the takedown of their content.	User-to-user and search services
Reporting online CSEA: If the platform is a UK platform or is a non-UK platform that does not already report, to report identified online CSEA to the NCA.	User-to-user and search services
Transparency reporting: to publish reports containing information about the steps they are taking to tackle online harm on those services.	Category 1, 2A (search) and 2B (user-to-user) services
Fraudulent advertising duty: to minimise the publication and/or hosting of fraudulent advertising	Category 1, 2A (search) and 2B (user-to-user) services
User verification and user empowerment duties: to offer optional user identity verification and user empowerment tools to give users more control over their online experience	Category 1 services
Freedom of expression and privacy: to assess the impact their policies have on users' free speech and privacy.	Category 1 services
Protected content: to put systems and processes in place to protect journalistic content and content of democratic importance.	Category 1 services

Duty	Services in scope[footnote 9]
Standalone pornography provision: to prevent children from accessing published pornographic content.	pornography publishers

Harms in scope of the Bill

The list below gives an indication of the online content or activity that we are interested in, as outlined in the Online Safety Bill.

Illegal content and activity that platforms will need to tackle

Priority illegal content is content that amounts to an offence on a list of the most harmful offences which companies must take proactive measures to prevent individuals from encountering via their service and to minimise the length of time for which any such content is present. These offences include:

Terrorism offences

- Child sexual exploitation and abuse offences
- Encouraging or assisting suicide
- Offences relating to sexual images i.e. revenge and extreme pornography
- Incitement to and threats of violence
- Hate crime
- Public order offences - harassment and stalking
- Drug-related offences
- Weapons / firearms offences
- Fraud and financial crime
- Money laundering
- Exploiting prostitutes for gain
- Assisting illegal immigration

The full list of priority offences can be found in Schedules 5 (terrorism), 6 (child sexual abuse and exploitation) and 7 (priority offences) of the Online Safety Bill. (https://bills.parliament.uk/publications/49376/documents/2822) Note: the government has also committed to adding coercive or controlling behaviour, Section 24 of the Immigration Act 1971 and Section 2 of the Modern Slavery Act, and the new Foreign Interference Offence (legislated for in the National Security Bill) to this list.

Companies also have to remove any other illegal content where there is an individual victim, where it is flagged to them by users or they become aware of it through any other means.

Content that is harmful to children

Companies will need to take specific action to prevent children from encountering content that has been designated as 'primary priority' harmful content to children, and must take an age-appropriate approach to protecting children from 'priority' harmful content. The indicative categories of content are:

Primary priority content (children must be prevented from encountering altogether):

- Pornography
- Content promoting self-harm (with some content which may be designated as priority content, e.g. content focused on recovery from self-harm)

Content promoting eating disorders (with some content which may be designated as priority content, e.g. content focused on recovery from an eating disorder)

- Legal suicide content (with some content which may be designated as priority content, e.g. content focused on recovery)

Priority content (companies need to ensure content is age appropriate for their child users):

- Online abuse, cyberbullying and harassment
- Harmful health content (including health and vaccine misinformation and disinformation)
- Content depicting or encouraging violence

Endnotes

---

1. Online Safety Bill impact assessment (https://www.gov.uk/government/publications/online-safety-bill-supporting-documents)
2. A full list of these changes can be found in published guidance: Online Safety Bill: government amendments at Lords report stage (https://www.gov.uk/government/publications/online-safety-bill-government-

amendments-at-lords-report-stage/online-safety-bill-government-amendments-at-lords-report-stage)

3. Online Safety Roadmap, Ofcom (https://www.ofcom.org.uk/__data/assets/pdf_file/0016/240442/online-safety-roadmap.pdf)
4. Note that pricing will vary materially depending on the type of age assurance solutions used and the volume as discounts would likely apply. For example, some AA providers offer solutions ranging from £0.01 per transaction.
5. It's important to note that the regulation applies to all in-scope platforms. This is a hypothetical cost on non- compliance, but was included as it was what several platforms reported.
6. Note that pricing will vary materially depending on the type of age assurance solutions used and the volume as discounts would likely apply. For example, some AA providers offer solutions ranging from £0.01 per transaction.
7. It is important to note that the regulation sets out to apply consistently across organisations so this cost should not be one platforms will have to bear, but is nonetheless a cost platforms were concerned would apply.
8. Additional information can be found in the Explanatory Notes of the Online Safety Bill: https://bills.parliament.uk/publications/49377/documents/2735
9. Thresholds for different categories of regulated services will be set out in secondary legislation; however, they will relate to a platform's number of users, functionalities and other relevant characteristics.