Online Harms Research: In-Scope Organisations' Approaches to Preventing Online Harm

2022

REVEALING REALITY

About Revealing Reality

Revealing Reality is an independent social research agency, working with regulators, government. Businesses, and charities to provide insight across sectors to inform decision-making, policy recommendations, and service design.

Studying how the digital world is shaping people's behaviours is something we do every day. We conduct detailed qualitative and quantitative behavioural research, observing how people really use digital products, services, and technology. This includes exploring how digital design shapes behaviour – across technology, gambling, financial products, health services, and more.

Visit www.revealingreality.co.uk to find out more about our work or to get in touch.

Contents

Executive summary

Project aim and objectives

The Department for Digital, Culture, Media & Sport (DCMS) and the Home Office jointly published the Online Harms White Paper (OHWP) in April 2019, setting out government's proposals for regulation and policies to tackle harms taking place online.

This research was commissioned to understand the potential impact of proposed regulation in two ways:

- exploring **how many** UK organisations could fall into scope of potential regulation
- assessing the **existing capabilities** and **approaches** to preventing online harm by organisations who do fall into scope, and **how regulation might affect this**

The findings of this research, along with the responses to the consultation and DCMS' stakeholder engagement, will enable DCMS to ensure regulation is realistic, achievable, and proportionate.

This research was carried out between August 2019 and February 2020. This report contains updated figures following re-analysis of the data to include new exemptions in Autumn 2020. Full details of this additional analysis and the impacts on the estimates can be found in the 'Revised figures - Autumn 2020 update' section on page 7.

Note: organisations who are currently expected to fall 'in scope of regulation' under the current interpretation of the OHWP and exemptions are also referred to as 'in-scope organisations' in this report. This includes any organisation who has a website or app which enables access to user-generated content (UGC) and peer-to-peer interaction (P2P).

The objectives were addressed in four phases, using quantitative and qualitative approaches.

Phase 1 What would bring an organisation into scope of regulation? *Developing a draft framework*	Phase 2 How many UK organisations fall into scope? *Random sample*	Phase 3 What types of organisations fall into scope and how do we categorise them? *Testing and refining the framework*	Phase 4 How do organisations currently tackle online harm & how would regulation affect this? *Interviewing organisations*
Objective: Create a clear framework for assessing whether or not an organisation is 'in scope' and develop a draft framework for categorising organisations	Objective: Provide an estimate of the number of organisations in the UK that would fall into scope of regulation	Objective: Test and refine the framework, to a) identify priority organisations for interviewing in Phase 4 and b) provide more detailed estimates of how many of *each type* of organization in the UK are in scope	Objective: Understand existing capabilities of in scope organisations to tackling online harm, and explore what the impact of potential regulation would be

Key findings from random sampling

Total figures suggested approximately **21,500 organisations, or approximately 0.3-0.4% of UK enterprises would be in-scope** (with a minimum of <1,000 and maximum of ~135,000), defined as organisations with apps or websites containing features that would bring them into scope and not meeting the criteria to be considered exempt.

For reference, original estimates before exemptions were included in the analysis in Autumn 2020 were that there were approximately **160-170,000 organisations with in-scope features** (with a minimum of <30,000 and a maximum of c.300,000). This equates to approximately 3% of UK enterprises.

Key findings from interviews with in-scope organisations

Most organisations who had significant amounts of UGC and P2P interaction on their platforms had many mitigations in place to prevent online harm

- Most of those with significant amounts of UGC and P2P interaction on their platforms, and who are therefore at the higher potential risk of online harm, already had many **mitigations** in place to protect users from harm:
    - All had human and automated moderation in place to varying degrees
    - All had reporting functions and procedures for users who experienced or witnessed harm
    - All had community guidelines in place to encourage limitation of harm
    - A number had implemented specific tools, namely PhotoDNA and GIFCT, to address the specific and illegal harms of Child Sexual Exploitation and Abuse (CSEA) and terrorist content
- Organisations responded to different harms in different ways – with most prioritising the identification and dealing with illegal harms

The costs and resources dedicated to mitigating online harm were generally proportionate to the organisation's risk of potential online harm

- Organisations who had greater amounts of UGC & P2P interaction on their platforms tended to have specific teams whose entire role was devoted to content moderation, with several platforms estimating that 15% of their total workforce work on content moderation specifically
- Estimates for the overall cost of protecting users from online harm were also higher for organisations where UGC & P2P interaction was primary to their platform. For example, a social media organisation estimated 15% of revenue was spent on 'user safety', a forum estimated 7% of their total annual expenditure on content moderation alone, and a messaging platform reported 10-15% of their annual budget was dedicated to 'user safety'
- Meanwhile, organisations where UGC and P2P interaction were secondary to their functioning (e.g. a retail platform with a reviews function) were less likely to dedicate significant proportions of their expenditure on protecting users from online harm. For example, one small (10-49 employees) retail site estimated it cost £50 per month to moderate reviews on their website, while others described the cost as 'negligible'

Organisations were motivated to invest in protecting their users for a variety of reasons, and many expected investment to increase over time

- There were a variety of motivations for investment in protecting users from harm, including creating a good user environment and fulfilling user expectations, as well as meeting the demands of advertiser and third-party suppliers, such as payment providers
- Measures that organisations took also depended, to a large extent, on corporate ideology—for example, a minority of organisations had entrenched ideas about the privacy of their users and claimed to collect as little data about them as possible. In comparison, those who were more successful (in economic terms) relied more heavily on an advertising sales-based business model which requires that they collect data on their users. This ideology affected the organisations' readiness for, and acceptance of, the OHWP
- Most organisations expected expenditure on protecting users to increase in order to remain competitive in the market and improve their platforms.
- Many said that if any additional activities were required to protect users from online harm due to new regulation, they were likely to divert existing budget for protecting users against harm, rather than

increase expenditure. Some organisations raised the concern that this might mean they would have to spend money doing things that might not necessarily be the most appropriate way for their organisation to prevent harm

- Overtly child-focused services were also keen to provide a safe and desirable environment for their users, and many had considered this when building their platforms

Overall, most organisations were supportive of the need for some form of regulation and were confident that under a reasonable interpretation of the OHWP they would not have to make major changes. However, some challenges were raised

- The majority of organisations felt that the measures they had in place to address online harms were sufficient for regulation under their interpretation of the OHWP, and were supportive of the need for regulation of some kind
- The challenges for most of the organisations were largely around **interpretation of more subjective 'harms'** such as bullying and misinformation, rather than illegal harms such as CSEA and terrorist content. Similarly, many struggled to reliably detect which of their users were children, and indeed, most organisations claimed they would adopt **age-gating**—a way to technically prevent children from accessing their site—without hesitation if a reliable service became available
- Organisations wished to know exactly what was required of them in particular circumstances in order to alleviate risk to themselves, while **maintaining flexibility to implement measures** that they considered to be sufficient and appropriate to the risks on their platform. This highlights a potential tension
- Some were concerned about the difference between **responsibility and liability**, believing that directors of organisations should not be liable for harms on their platform, but should take responsibility for preventing such harms
- Few of those with significant UGC and P2P interaction already **produced public transparency reports**, and some claimed to lack the data to do so. Others raised the costs of legal advisers and consultants to prepare this. Smaller organisations with less UGC and P2P interaction on their platforms also did not produce transparency reports, but considered that this would be feasible

Introduction

Background & project aims

Exploring the impact of the proposed online harms regulation

The Department for Digital, Culture, Media & Sport (DCMS) and the Home Office jointly published the Online Harms White Paper (OHWP) in April 2019, setting out government's proposals for regulation and policies to tackle harms taking place online.

This research was commissioned to understand the potential impact of proposed regulation in two ways:

- exploring **how many** UK organisations could fall into scope of potential regulation
- assessing the **existing capabilities** and **approaches** to preventing online harm by organisations who do fall into scope, and how regulation might affect this

The findings of this research, along with the responses to the consultation and DCMS' stakeholder engagement, will enable DCMS to ensure regulation is realistic, achievable, and proportionate.

This research was carried out between August 2019 and February 2020.

Note: organisations who are currently expected to fall 'in scope of regulation' under the current interpretation of the OHWP are also referred to as 'in-scope organisations' in this report.

Revised figures - Autumn 2020 update

In response to stakeholder consultation following the initial publication of the Online Harms White Paper, a number of additional exemptions (see below) were incorporated into DCMS's impact evaluation. This report contains updated estimates based on including these exemptions in our analysis. Both the original and updated estimates/data are shown for clarity.

Overall, the impact of the exemptions is a significant reduction in the number of organisations we would expect to fall in-scope of regulation.

The effect of the new exemptions appeared to be to increase the threshold at which point an organisation becomes in-scope, excluding many types of organisations who were classified as being in the lower tiers originally—i.e. those who were technically in-scope, but with limited peer-to-peer functionality/access to user-generated content.

Following the inclusion of the new exemptions, those classifying as in-scope are largely the types of organisations who were already classified in the higher tiers originally, for whom peer-to-peer interaction and enabling access to user generated content are a more important part of their core business functions/services/ products.

The original research found that under a broad interpretation of the OHWP, around 3% of all UK businesses could be considered in scope, equating to approximately 180,000 businesses. Inclusion of the below exemptions reduced this figure to approximately 21,500, or ~0.3 - 0.4% of UK businesses. Please note, this includes the addition of some logical extra estimates of organisations we know to be in-scope organisations not present within the IDBR sample (e.g. adding in dating sites).

Exemptions from regulation

Below is a description of the exemptions that were confirmed after the initial research was conducted, and which are accounted for in the revised figures/estimates presented in this report (alongside the original figures).

- **'Low risk functionality' exemption**: The Online Safety Bill will exempt user comments on digital content provided that they are in relation to content directly published by a platform/service. This will include reviews and comments on products and services directly delivered by a business, as well as 'below the line comments' on articles and blogs.

- **Services used internally by businesses**: This is defined as a service (or distinct part of a service), managed by an organisation, whose primary purpose is to host members' user-generated content and enable interactions between members within that organisation. This encompasses online services which are used internally by organisations such as intranets, customer relationship management systems, enterprise cloud storage, productivity tools and enterprise conferencing software.
- **Network infrastructure**: Any service which does not have direct control over the User Generated Content on their platform. In practice, this removes network infrastructure such as Internet Service Providers, Virtual Private Networks and content delivery services as they do not have any control over an individual piece of content. This exemption also applies to business-to-business services e.g. white label or software as a service (SaaS) services offered to businesses where, again, the supplier does not have control over specific pieces of content or activity.
- **Educational platforms**: Online services managed by educational institutions, including early years, schools, and further and higher education providers. This includes platforms used by teachers, students, parents and alumni to communicate and collaborate. This includes platforms like intranets and cloud storage systems, but also "EdTech" platforms.
- **Email and telephony**: Email communication, voice-only calls and SMS/MMS remain outside the scope of legislation.

Furthermore, business-to-customer interactions are not considered user generated content and will also be out of scope (for example video and email interactions between a user and a business). An example of this would be a complaints box where users can interact with a business as well as patient-doctor virtual services where users can have a virtual appointment with a physician.

Research approach

A four phase, mixed-methods approach

Phase 1 What would bring an organisation into scope of regulation? *Developing a draft framework*	Phase 2 How many UK organisations fall into scope? *Random sample*	Phase 3 What types of organisations fall into scope and how do we categorise them? *Testing and refining the framework*	Phase 4 How do organisations currently tackle online harm & how would regulation affect this? *Interviewing organisations*
Objective: Create a clear framework for assessing whether or not an organisation is 'in scope' and develop a draft framework for categorising organisations	Objective: Provide an estimate of the number of organisations in the UK that would fall into scope of regulation	Objective: Test and refine the framework, to a) identify priority organisations for interviewing in Phase 4 and b) provide more detailed estimates of how many of *each type* of organization in the UK are in scope	Objective: Understand existing capabilities of in scope organisations to tackling online harm, and explore what the impact of potential regulation would be

The research consisted of four phases, each with a specific research objective:

The research involved a mixture of research methods. Phase 1 and 3 included expert interviews and desk research; an audit of a random sample of UK enterprises was used to estimate the number of in-scope organisations in Phase 2; and qualitative interviews with in-scope organisations were undertaken in Phase 4. For further detail on the methods used in each phase see the relevant section.

Phase 1: What would bring an organisation into scope of regulation?

Framework development

Phase 1 objectives

Create a clear framework for determining whether or not any given organisation would be 'in scope' of potential regulation, as currently stated within the OHWP, and identify different ways of categorising these 'in scope' organisations.

This would enable us to:

- a) objectively assess, using a consistent range of measures, whether or not an organisation was 'in scope'. This was required ahead of Phase 2, in which organisations were assessed as to whether or not they were in scope
- b) decide which organisations to speak to in Phase 4 – ensuring the research covered the full range of organisations who may fall into scope of regulation. This would enable us to explore how different factors, such as business size, may affect their capability to tackle online harm or comply with potential regulation

Defining 'in scope'

To determine whether or not an organisation is 'in scope' of the potential regulation proposed in the OHWP, it is necessary to clearly define what would bring an organisation into scope. As per the OHWP, the proposed regulatory framework would apply to organisations who "allow users to share or discover user-generated content or interact with each other online".

'In-scope' for the purposes of this research was therefore determined by whether a website or app:

- enabled peer-to-peer (P2P) interaction – i.e. allowing users to interact with other users in any way
- *and/or* enabled access to user-generated content (UGC) – i.e. being able to see content of any kind that was created and uploaded by another user
- *and* was not excluded in line with additional exemptions confirmed in Autumn 2020 (see Exemptions on page 7)

If these criteria are met, then an app or site is considered as being in scope for potential regulation. Importantly, this does not mean that the organisation in question would be affected by regulation, but rather simply that it could be, depending on what parameters or definitions any future regulation employs. The additional exemptions have the effect of removing from scope some organisations who previously would have been technically in-scope but who were realistically very unlikely to be affected by any regulation.

Method

The Organisation Categorisation Framework (OCF) was developed using extensive **desk research**, interviews with **experts** in this area—such as the IWF, Childnet, and Internet Matters—and DCMS's **pre-existing knowledge**. It was drafted in Phase 1 at the start of the project and then further tested and refined in Phase 3, using organisations identified the random sampling in Phase 2.

Developing the OCF

The OCF was drafted by first identifying all of the factors that define whether or not an organisation is in scope and any factors that could affect its ability to tackle online harm. This produced a long list of factors that was broadly split into three areas:

- Features
- Mitigations
- Organisational factors

This section will provide further detail on each of these areas.

Features

To determine whether or not an organisation allowed users to share or discover user-generated content or enable peer-to-peer interaction, we first looked at the technical **'features'** of sites and apps that enable this behaviour.

Features were separated into eight themes and the table below illustrates the features identified as part of this research. This is not necessarily an exhaustive list and should be seen as something that can be added to over time as new features are identified, created or developed. However, it is intended to be a relatively complete list of all current features currently used by in-scope organisations.

The majority of features can be identified by using a site or app and exploring the features available as a user. Certain platforms require certain permissions or log in details to access all of their features.

Table 1 showing features of platforms that enable users to share or discover user-generated content or enable peer-to-peer interaction, split across eight themes

Theme	In-scope feature	Explanation, if applicable
Posting *Uploading or broadcasting your own content*	• Posting UGC of any kind	
	• Livestreaming	
	• Sharing live location	
	• Time-limited sharing	Posts that 'disappear' after a certain amount of time or after viewing
	• Screen sharing	Showing others what is presented on the screen while using a device
Sharing *Sharing content that already exists on a platform*	• Sharing content that exists on the platform through a sharing function	Using a link to share one's own or somebody else's content, including content posted by the site, with others within the platform in a similar manner to own posts
	• Sharing content that exists on the platform through personal messages	Using a link to share content as above with another user via a messaging service hosted by the platform
	• Sharing content that exists on the platform through group messages	As above, but to multiple users either in the same message or at same moment
	• Sharing files hosted online	Inviting others to view and/or edit files hosted online, for example through Google Drive
Reacting *Reacting to content*	• Liking content	
	• Disliking content	
	• Up-voting content	Similar to liking, but directly promotes the exposure of the content to others

	• Down-voting content	Similar to disliking, but directly reduces the exposure of the content to others
	• Reacting to content in other ways	For example, leaving a star rating
	• Liking comments	
	• Disliking comments	
	• Up-voting comments	
	• Down-voting comments	
	• Reacting to comments in other ways	For example, adding particular emojis to the comments
	• Liking messages	
	• Disliking messages	
	• Up-voting messages	
	• Down-voting messages	
Messaging *Sending messages to others*	• Personal message	
	• Group message	
	• Messaging only approved contacts	Requires a request and acceptance feature
	• Messaging anybody whose information you have	Requires you to have specific information to message individuals
	• Messaging anybody on the platform	Allows messaging to anybody, potentially through intentional linking with strangers
Calling *Video or voice calling*	• Voice calling anybody on the platform	
	• Voice calling contacts only	
	• Video calling anybody on the platform	
	• Video calling contacts only	
Commenting *Commenting on content*	• Posting comment under content posted by the host of site	For example, a review of a product
	• Posting comments on user posts	
	• Commenting on comments	For example, being able to reply to specific comments rather than simply posting below
Tagging *Creating links between bits of content and people or places*	• Tagging people or groups in posted content	Where tagging identifies a specific individual or group and notifies them
	• Tagging people or groups in comments	
	• Tagging people or groups in messaging	
	• Geo-tagging	Tagging posts with locations but, unlike sharing live location, these are static
Discovering *Enabling users to discover content*	Search function linking to UGC	
	Display or feed of UGC	

Mitigations

Mitigations were defined as 'anything that makes a platform or feature less potentially risky to a given user'. These were broken down into categories such as:

Table 2 showing mitigations to make platforms less potentially risky to users, with examples

Mitigation	Examples
Moderation - **When**	**Pre-moderation**: Submitted content will be put in a queue and checked by a moderator before being made visible
	Post-moderation: All content is displayed on the site immediately and moderation is done after, often prompted by reports from users
Moderation - **How**	Automated moderation
	Human moderator - Employed
	User moderators/editors - Volunteer
Moderation – **What**	Editing content
	Removing content
	Removing/banning and suspending users
Reporting functions	User reporting content to site
	Flagging content e.g. trigger warnings or NSFW (not safe for work) warnings done by users
Parental control	Parental controls
Access to site	1) Can you access the site without logging in, agreeing to anything, or ticking any boxes? Which features can be accessed without doing these things? 2) How do you make a user? What information is required of you and do you need to verify this information? E.g. Age verification, email or text verification, log in with CAPTCHA 3) What features can only be accessed through making a user?
Platform support and advice	Community guidelines
	Signposting to services/support
	Automatic deterring messages when certain words/phrases are typed

There are a few limitations to using mitigations to categorise organisations:

- It is hard to identify all the mitigations a platform has in place without first speaking to the organisation. Unlike features, some mitigations are not obvious from looking at or using their platform. For example, it is not obvious when or how an organisation is using automated moderation from simply observation of their platform—you can't always know what, if anything, your post is being scanned for. Moreover, different organisations may use the same mitigations, but in different way—for example, different rules for blocking users.
- It is also difficult to assess how effective these are in mitigating online harm.

Organisational factors

A number of factors relating to the organisations as a whole, as opposed to the sites or apps they operate, could also influence the potential for risk, as well as their approaches to dealing with online harm and attitudes towards potential regulation.

These included:

Table 3 showing organisational features which may impact how they operate, potential for risk and approaches to dealing with online harm

Factor	Comment
Location of organisation	Organisations had to have a UK presence beyond access to the website. For example, US-based platforms with no UK subsidiary or office could not be affected by UK regulation.
Host of platform	Organisations who hosted their own website would be responsible for its content, while those who outsourced this or used an intermediary would not.
Type of platform	Examples: social media, image sites, blogs, dating platforms, e-commerce sites and business-to-business services.
Type of business	Including but not limited to charities, limited companies, co-operatives and educational body.
Whether UGC and peer-to-peer interaction was primary or secondary to the organisation's purpose	Could affect the likelihood of harms and extent of mitigations in place, as well as their awareness of the OHWP
Vulnerability of user	For example, platforms aimed at children or with particular vulnerabilities, such as domestic abuse survivors
Size of organisation	Number of employees
Number of users	
Financial factors	Turnover, amount spent on protecting users from online harm

As with the mitigations, there were limitations to using all of the organisational factors identified to categorise organisations. For example, some of these factors were not obvious from researching the platform, such as the specific details on what organisations spend money on.

Identifying the most important factors of the OCF

Having considered all the features, mitigations and organisational factors, the choices for incorporation into the OCF were pragmatic. To use a consistent framework for assessing organisations, the information needed to categorise organisations available had to be readily accessible for each organisation.

The two primary categorisation criteria we chose to incorporate into the OCF were therefore those that could be *objectively assessed* based on readily available information about the platforms:

- **Features**: these could often be assessed by viewing the platform or creating an account, though some organisations restricted access to some features from public users—for example, schools only allowed access to students, parents, and teachers via portals.

  The number and type of features a platform enabled could be used a proxy for the potential risk of online harm. This assumed that each feature comes with a potential to enable online harm in some way. Given that certain features, such as 'liking a comment' have an inherently lower potential to enable online harm compared to a feature like 'livestreaming', the type of feature also had to be accounted for.

- **Business size**: using the proxy of number of employees, categorisation by business size could be achieved through a number of public sources.

  Business size was important due to the OHWP's focus on *proportionately* – ensuring that processes to prevent online harm were proportionate to the capability of an organisation. Business size also tended to correlate with the number of users a platform had, though user numbers were often hard to verify without speaking to organisations.

Employee number was used as a first method of categorising organisations, as described in Phase 2, while features were incorporated into the classification of risk, or 'tier', as described in Phase 3.

A proxy for the *potential* risk of online harm was required as there is no consistent, publicly available way to assess the online harm *actually* enabled by different organisations.

Phase 2: How many UK businesses fall into scope?

Random sample

Phase 2 objectives

To provide an **estimate of the number of organisations** in the UK that would fall into scope of potential regulation.

This was achieved through the testing of a random sample of UK enterprises and extrapolating the results to the whole population of UK enterprises. These figures provided an estimate for the total number of organisations in the UK who might be considered in scope for potential regulation.

IDBR Sampling Approach

What is the IDBR?

The Inter-Departmental Business Register (IDBR) is a comprehensive list of UK businesses used by government for statistical purposes. It covers 2.6 million businesses in all sectors of the UK economy, other than very small businesses (those without employees and with turnover below the tax threshold) and some non-profit-making organisations.

The two main sources of input were the Value Added Tax (VAT) system from HMCR (Customs) and Pay As You Earn (PAYE) from HMRC (Revenue). Additional input came from Companies House, Dun and Bradstreet, and our business surveys.

Disproportionate stratified random sample

The IDBR was divided into groups, or 'strata', based on organisation size (number of employees). A disproportionate stratified random sample of 500 organisations was taken, comprised of 100 randomly selected organisations in each of the five size categories. The table below shows this division.

Table 4 showing size of organisation by number of employees across the IDBR sample

Organisation size (by no. of employees)	**N =**
0, or sole traders	100
1 – 9	100
10 – 49	100
50 – 249	100
250 +	100
Total	***500***

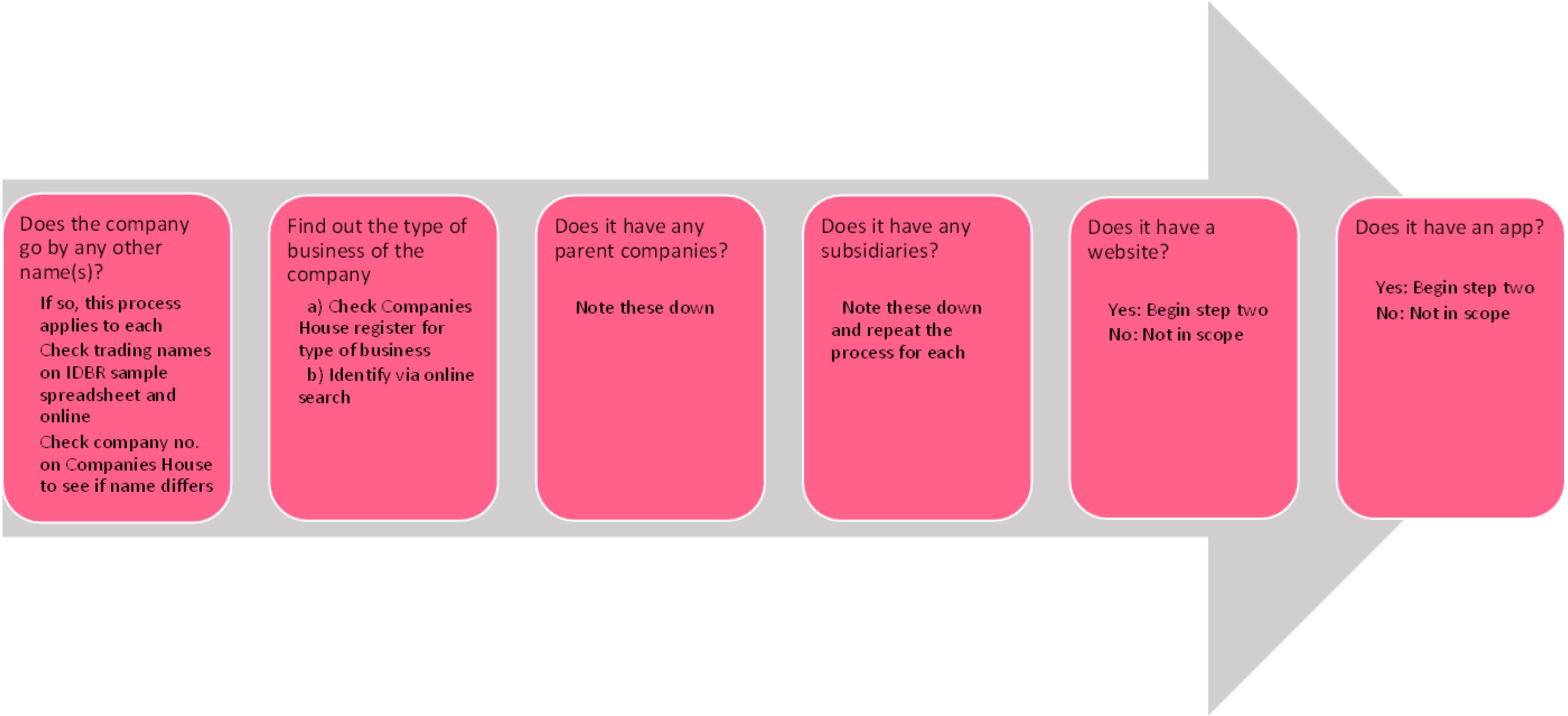
A key advantage to this approach was to guarantee a minimum number of organisations (100) to be tested within each size group. If a purely random sample had been taken, we would have had only a few large organisations—not nearly enough for any reasonable analysis. For analysis and estimates within each size group the five samples were treated independently; for total population estimates the samples were weighted back to their natural proportions.

A total sample size of n=500 was large enough to provide robust estimates for the number of in-scope organisations as it ensured a relatively small margin of error at the 95% confidence level (between ±2.6 to 4.4 percentage points). This sample size was also manageable—every organisation within the sample had to be manually reviewed and categorised to determine what in-scope features and mitigation practices they had in place on their website or app (if they had one) was possible. For each group (or 'strata') of n=100 organisations, the margin of error at the 95% confidence level varied between ±2.2 and 6.8 percentage points, as displayed below at Stage 3.

Testing sample organisations

By using the eight themes (posting, sharing, messaging, calling, reacting, commenting, tagging and discovering) and their sub-features as set out in the OCF framework, each of the 500 organisations in the sample were assessed through a two-stage process to determine whether any in-scope features were present.

Step One



Does the company go by any other name(s)?

If so, this process applies to each

Check trading names on IDBR sample spreadsheet and online

Check company no. on Companies House to see if name differs

Find out the type of business of the company

a) Check Companies House register for type of business

b) Identify via online search

Does it have any parent companies?

Note these down

Does it have any subsidiaries?

Note these down and repeat the process for each

Does it have a website?

Yes: Begin step two

No: Not in scope

Does it have an app?

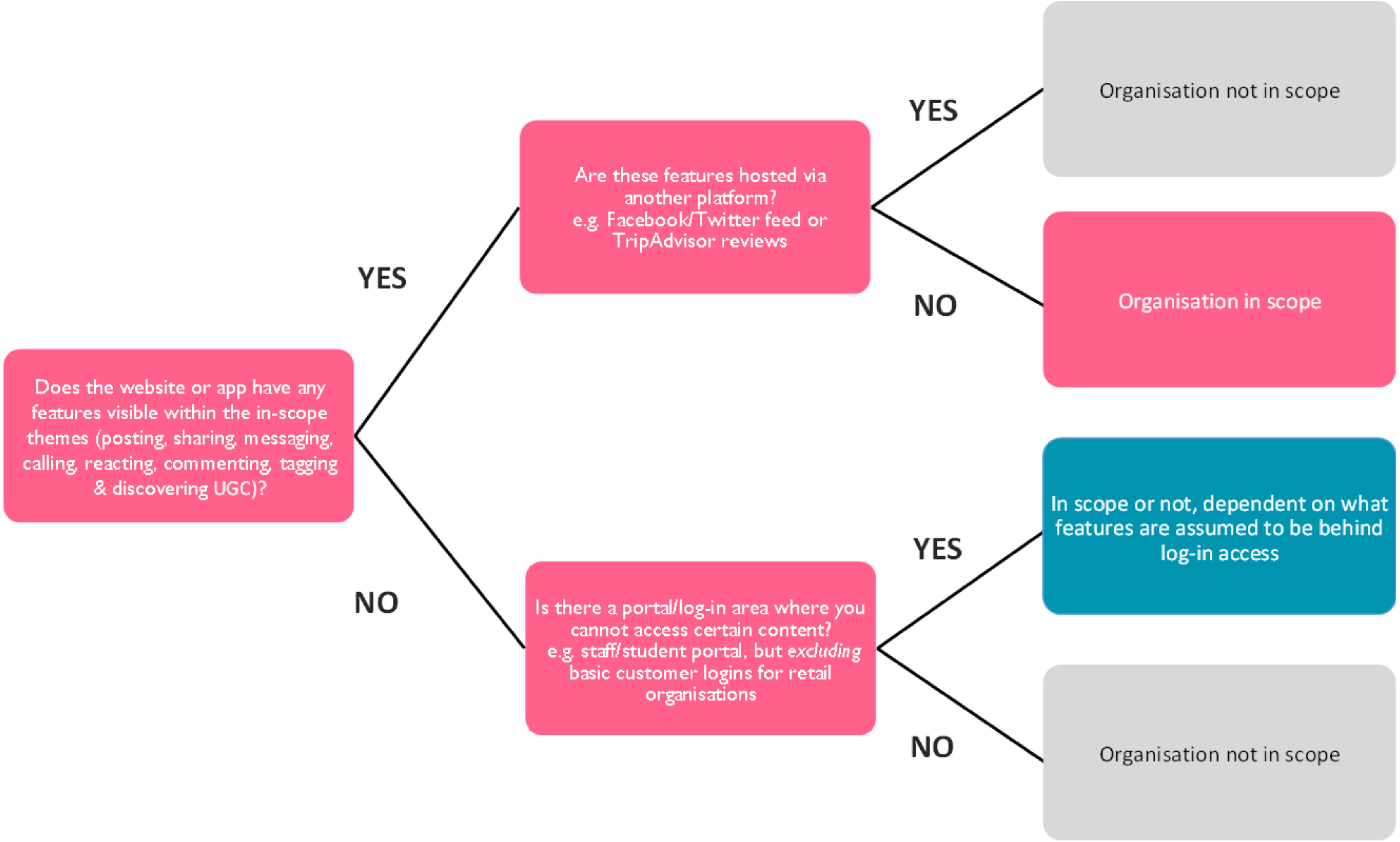
Yes: Begin step two

No: Not in scope

The diagram below shows the first stage of the assessment process. This step ensured that all websites or apps under a sample company's responsibility were considered and given an initial assessment of whether the organisation might be in scope.

Step Two

Having identified the websites or apps relating to each organisation in the sample, we proceeded to assess the websites or apps against the features in the OCF framework. The process diagram below, starting from the left, shows the assessments made to determine whether an organisation was in scope.


Does the website or app have any features visible within the in-scope themes (posting, sharing, messaging, calling, reacting, commenting, tagging & discovering UGC)?
YES
Are these features hosted via another platform? e.g. Facebook/Twitter feed or TripAdvisor reviews
YES
Organisation not in scope
NO
Organisation in scope
NO
Is there a portal/log-in area where you cannot access certain content? e.g. staff/student portal, but *excluding* basic customer logins for retail organisations
YES
In scope or not, dependent on what features are assumed to be behind log-in access
NO
Organisation not in scope

Estimating the number of in-scope organisations

Having assessed the organisations in the sample, the percentages of organisations found to be in scope were used to estimate the total number of in-scope organisations. This number refers to the organisations registered in the UK, which may or may not have a website, and *not* to the websites available to be accessed in the UK.

The number of in-scope organisations varied by the company size, with larger organisations more likely to have a site or app with in-scope features. As the IDBR is largely representative of all UK enterprises, the following percentages of in-scope organisations are considered to refer to the whole UK economy. The tables below give the breakdown of organisations found to be in scope within each strata for the total sample and for private sector enterprises only.

Table 5 showing estimated percentage of in-scope organisations by size, including Autumn 2020 exemptions

Revised figures, including Autumn 2020 exemptions			
Total (private, public and third sector)			
Size group	**Number of employees**	**Percentage in scope (within strata)**	NB. Once exemptions included sample too small to split into private sector only
Micro[1]	0 – 9	0.3%	
Small	10 – 49	0.0%	
Medium	50 – 249	2.0%	
Large	250+	5.0%	

Table 6 showing estimated percentage of in-scope organisations by size, pre-exemptions

Original figures, pre-exemptions						
Total (private, public and third sector)				**Private sector enterprises only**		
Size group	**Number of employees**	**Percentage in scope (within strata)**		**Size group**	**Number of employees**	**Percentage in scope (within strata)**
Micro[2]	0 – 9	2.6%		**Micro**	0 – 9	2.59%
Small	10 – 49	6.0%		**Small**	10 – 49	5.38%
Medium	50 – 249	9.0%		**Medium**	50 – 249	9.41%
Large	250+	14.0%		**Large**	250+	10.96%

How many enterprises within the whole economy are in scope?

Using the percentage of in-scope organisations within our sample, recent (2019) BEIS business population estimates we calculated the estimated number of in-scope enterprises within each size group. and maximum and minimum figures were calculated. The margin of error (MoE)—based on sample sizes used and number of in-scope organisations identified in each size group—enabled us to provide minimum and maximum estimates.

[1] Micro represents both the groups with 0 employees, or sole traders, and with 1 – 9 employees, and has been weighted to reflect their natural proportions among all UK organisations.

[2] Ibid

The table below gives the breakdown of the number of organisations estimated to be in scope within the whole economy, with MoE, and the potential range of in-scope organisations.

Table 7 showing number of organisations estimated to be in scope within the whole economy, with MoE, and the potential range of in-scope organisations, including Autumn 2020 exemptions

Revised figures, including Autumn 2020 exemptions								
		Whole economy		**MoE**	**Organisations in scope?**		**In-scope organisations**	
	Number of employees	Count within whole economy	Percentages within whole economy	MoE at 95%: +/-	Yes	No	Minimum	Maximum
Micro	0 – 9	5,689,935	95.33%	1.94%	17,070	5,672,865	0.6	127,455
Small	10 – 49	226,900	3.80%	1.95%	-	226,900	-	4,425
Medium	50 – 249	41,640	0.70%	2.74%	833	40,807	2	1,974
Large	250+	10,445	0.17%	3.82%	522	9,923	123	922
	Total	**5,968,920**	100.00%		**18,425**		126	134,775

Total figures suggested approximately **18,000 organisations with in-scope features** (with a minimum of <1,000 and a maximum of ~135,000). This equates to approximately 3% of UK enterprises.

Table 8 showing number of organisations estimated to be in scope within the whole economy, with MoE, and the potential range of in-scope organisations, pre-exemptions

Original figures, pre-exemptions								
		Whole economy		**MoE**	**Organisations in scope?**		**In-scope organisations**	
	Number of employees	Count within whole economy	Percentages within whole economy	MoE at 95%: +/-	Yes	No	Minimum	Maximum
Micro	0 – 9	5,689,935	95.33%	2.21%	148,126	5,541,809	22,606	273,646
Small	10 – 49	226,900	3.80%	4.65%	12,614	214,701	3,054	24,176
Medium	50 – 249	41,640	0.70%	5.60%	3,748	37,721	1,415	6,080
Large	250+	10,445	0.17%	6.77%	1,462	9,300	755	2,169
	Total	**5,968,920**	100.00%		**166,949**		**27,830**	**306,069**

Total figures suggested approximately **160-170,000 organisations with in-scope features** (with a minimum of <30,000 and a maximum of c.300,000). This equates to approximately 3% of UK enterprises.

Phase 3: What types of organisations fall into scope?

Testing and refining the framework

Phase 3 objectives

This phase involved testing and refining the OCF using the random sample data in order to:

- Identify potential priority organisations to interview in Phase 4
- Provide more detailed estimates of the numbers of each type of organisation in the UK that are in scope

Scoring and tiering organisations

Developing a way to score features

As discussed in Phase 1, the **number** and **type** of features a platform enables was used as a proxy for risk of potential online harm.

In order to develop a way to categorise organisations based on the potential for risk, we needed to develop a system to score an organisation's features.

Each feature was assigned a score, and the organisation's score was calculated from the sum of the scores of all of their features. This meant that the more features an organisation had, the higher their score.

Features that had an inherently greater potential online harm, such as livestreaming, received higher scores than features that were more limited, such as being able to like a comment.

The scores assigned to each feature were based on desk research and interviews with experts, which highlighted the features commonly associated with online harm. Nevertheless, the numerical values given to each feature were subjective, and arbitrary when taken alone. They only became useful and meaningful as a way to compare and categorise organisations across a consistent scale.

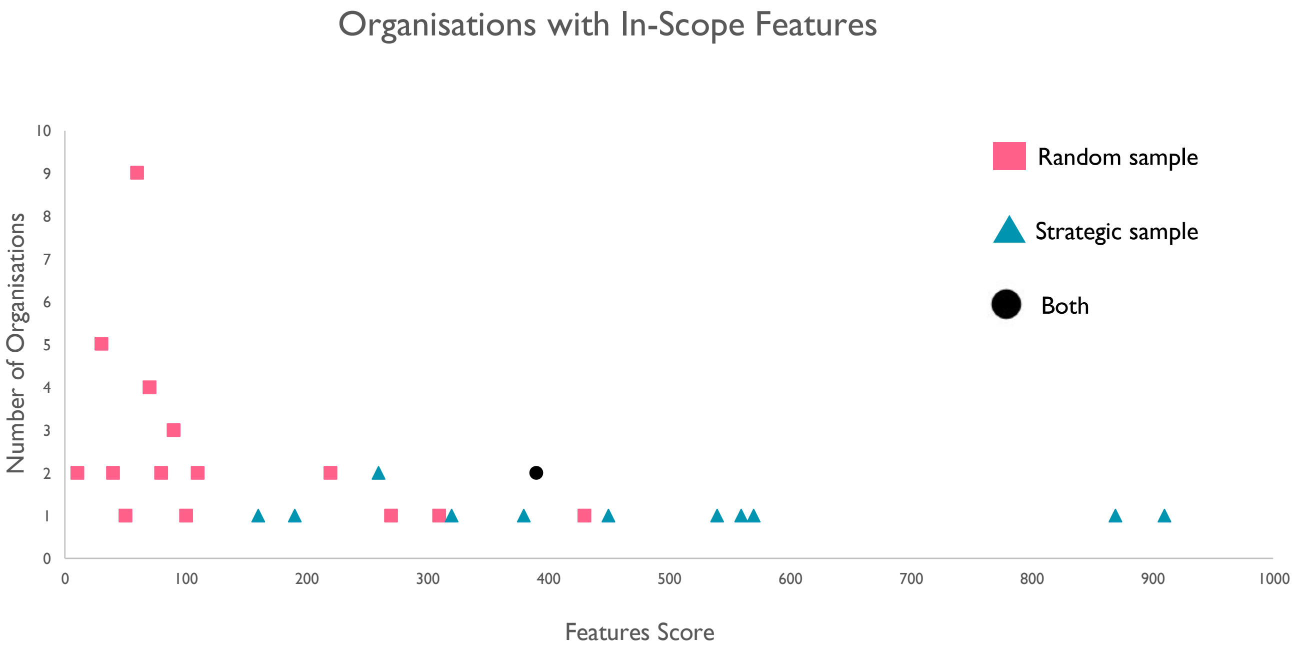
Again, it is important to remember that scores associated with features indicate the *potential* for online harm rather than inferring anything about their actual enabling of online harm.

Testing the scoring system with in-scope organisations

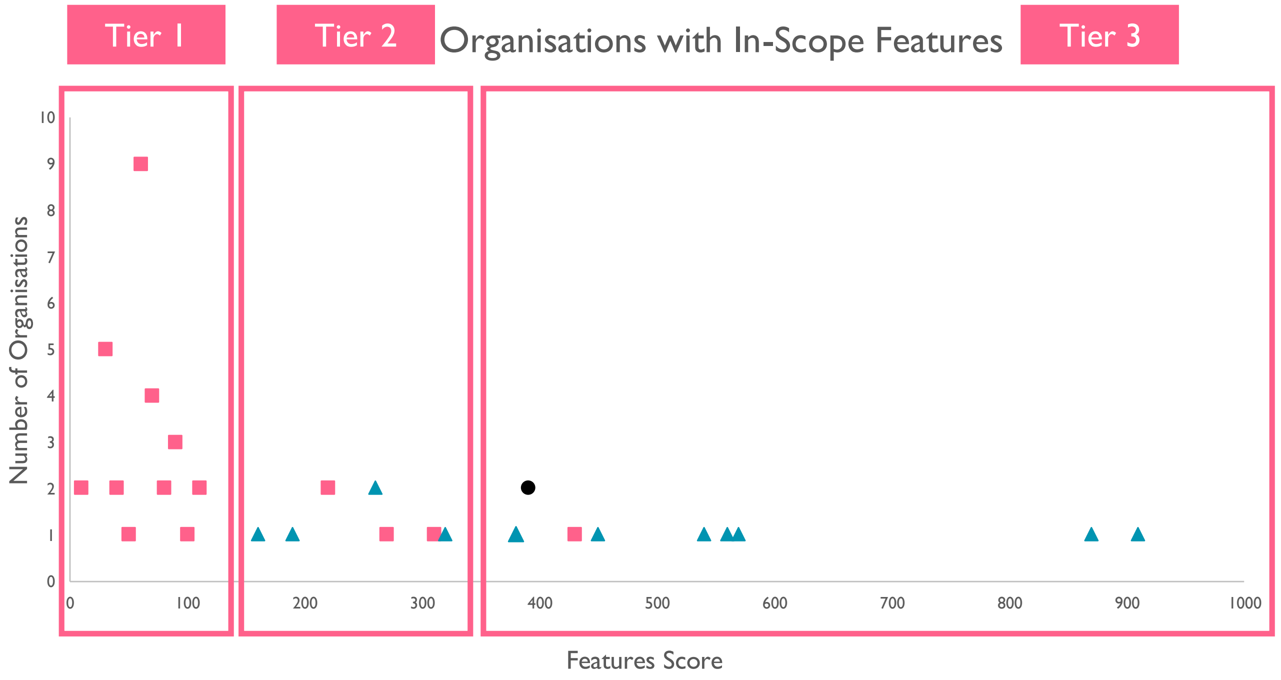
Organisations who were identified as in scope in the random sample were scored using this system. Similarly, some of the large social media organisations who were likely to have lots of in-scope features were scored, and the results were mapped on a graph for comparison.

Note: the highest possible score, if an organisation had all of the in-scope features, was 1000.

In the graph below, pink squares indicate organisations from the random sample, while the blue triangles indicate organisations from a variety of sectors, such as social media, forums or dating apps, who were selected for assessment. The black dot indicates that an organisation from the random sample and a specifically selected organisation both had a feature score of 390.



From contextual understanding of the type of platforms and the features they enabled, draft boundaries were drawn based on the features score. This produced three potential tiers for risk as grouped below. Tier 1 being the lowest potential risk of online harm and tier 3 being the highest. As can be seen in the table below, the majority of organisations in the random sample fell into Tier 1.



The table below gives an overview of the features typically seen within each of the tiers and the types of organisation running these websites or apps.

Table 8 showing features typically seen within each of the tiers and the types of organisation running these websites or apps

	Description	Example feature	Example type of organisation
Tier 1	Minimal in-scope features and mostly low-risk features, where main purpose of site/app tends not to be peer to peer interaction or user generated content Score: 1 - 149	• Comments section • Ability to like content	• Retail website • Personal or business blogs
Tier 2	A number of in-scope features Score: 150 - 349	• Ability to post content • Message somebody you 'know' or have 'friended'	• Forum • Dating app • Online gaming
Tier 3	Lots of in-scope features/high-risk in-scope features Score: 350+	• Feed of UGC • Live-streaming service • Ability to contact 'unknown' users	• Social media • Streaming service

One factor that wasn't accounted for in this model of potential risk of online harm was the **vulnerability of users.** For example, where platforms were clearly designed for and used by children, the risk of certain online harms is increased, and extra precautions may be necessary. We therefore decided to 'bump' organisations up a tier if children or vulnerable audiences were clearly identified as the target audience.

Creating 'tiers' of risk allowed us to differentiate between different organisations. This in turn enabled us to understand who to prioritise in Phase 4 in order to focus on organisations most likely to be affected by regulation.

Refining the random sample

Using the tiering system that was developed, a more detailed breakdown of the businesses operating in the United Kingdom could be produced.

Part 1: Tiered random sample estimates

Using the features scores and tiering for the in-scope organisations within the random sample, the organisations within the UK could be broken down further. More detailed estimates of the number of organisations within each organisation size group and tier were produced.

These are shown below. It is important to note that the number of in-scope organisations within the random sample was low. Following further division into tiers, the number of organisations within each was very low and all figures given should be treated with caution.

Estimated numbers are shown rounded up to the nearest thousand. In brackets is shown the range – the minimum and maximum estimates based on factoring in the margin of error for each sample size group. The main figures in each box represent the midpoint of these estimates.

Table 9 showing the estimated number of organisations in-scope within each organisation size group and tier, including exemptions

Revised figures, including Autumn 2020 exemptions					
	Total	**Micro**	**Small**	**Medium**	**Large**
Total in scope	**18.5k**	**17,000**	**0**	**<1000**	**<1000**
Tier 1	**<10k**	8,500	-	400+	400+
Tier 2	**<10k**	8,500	-	-	100+

Tier 3	**<1k**	-	-	400+	-

Table 10 showing the estimated number of organisations in-scope within each organisation size group and tier, pre-exemptions

Original figures, pre-exemptions					
	Total	**Micro**	**Small**	**Medium**	**Large**
Total in scope	**170k** (25 – 302k)	**150k** (20 – 270k)	**14k** (3 – 24k)	**5k** (1.5 – 6k)	**1.5k** (<1 – 2k)
Tier 1	**c.130k**	c.120k	c.14k	c.3.5k	c.1k
Tier 2	**c.28k**	c.28k	-	-	c.<1k
Tier 3	**c.<1k**	-	-	c.<1k	-

These estimates were drawn from the random sample. However, small tier 2 organisations, for example, are known to exist, as are large tier 3 organisations, though none were identified in the random sample.

Part 2: Adding known organisations and sectors

Acknowledging the gaps in the random sample analysed to reach the initial figures, we also included estimates for *types* of organisation we identified as likely to sit in certain tiers. For example, e-commerce or retail sites[3] that we analysed tended to have enough features to put them in Tier 2. The estimates in the table below include additional enterprises based on the assumption that all (or at least most) of relevant types of organisation should therefore be included. In this way we were able to fill some known gaps from the random sample. It is important to note that these additions do not represent an exhaustive list of all types of organisation that could be in-scope, but are an attempt to deal with some of the larger groups to provide a more realistic estimate. For example, these include crowdfunding or fundraising sites, dating sites and forums (assuming approximately an additional 3,000 orgs will split across small and medium sized organisations approximately 1:2, and evenly across tier 1 and tier 2).

The threshold for Tier 3 requires a wide range of features that only a small number of organisations would employ, for various reasons. Through extensive desk research searching for types of organisation (e.g. social media) and types of features (e.g. video calling ability) we identified as many Tier 3 organisations as possible, and used these to estimate the numbers shown below.

Table 11 showing the estimated number of organisations in-scope within each organisation size group and tier, including exemptions and additional enterprises

Revised figures, including Autumn 2020 exemptions							
	Total	**Micro**	**Small**	**Medium**	**Large**	**Total range**	**% of UK organisations**
Total in scope	**~21.5k**	**17,000**	**<3,000**	**<1,000**	**<1,000**	**<5k – 35k**	**0.3-0.4%** (0.01% – 1%)
Tier 1	**~10k**	8,500	~500*	~1,000*	400+	5-15k	**<0.5%**
Tier 2	**~10k**	8,500	~500*	~1,000*	100+	5-15k	**<0.5%**
Tier 3	**<1k**	50**	50**	400+	50*	<1k-2k	**<0.002%**

Please note: all figures are estimates based on the available data and should be treated as such. Minimum and maximum ranges have been included (in brackets) to account for margin of error.
Figures marked with a * are either wholly or partly made from estimates of known types of organisations who may be in-scope as explained above the table. Figures marked with ** indicate where additions were made to ensure the estimates contained a cautious overestimate – while no organisations that met the size and tier requirements were identified in the research, it is possible there are a small number in the whole population who do.
Figures without a * come from analysis of the random sample of UK enterprises and extrapolating the figures. In assessing the potential impact of regulation on UK enterprises we are working on the assumption that an overestimate is more cautious, and therefore appropriate, than underestimating the impact.

[3] Contact data for UK- based e-commerce sites

Table 12 showing the estimatednumber of organisations in-scope within each organisation size group and tier, pre-exemptions, including additional enterprises

Original figures, pre-exemptions							
	Total	**Micro**	**Small**	**Medium**	**Large**	**Total range**	**% of UK organisations**
Total in scope	**170k** (25 – 302k)	**150k** (20 – 270k)	**25k** (8 – 35k)	**<10k** (1.5 – 15k)	**<3k** (<1 – 3k)	**<200k** (25 – 300k+)	**3-4%** (0.5 – 5.5%)
Tier 1	**c.143.5k**	c.120k	c.16k	c.6.5k	c.1k	**130-150k**	**2%**
Tier 2	**c.36k**	c.28k	c.5k*	c.3k*	c.1k*	**20-40k**	**<1%**
Tier 3	**c.3k**	<10*	<10*	<50*	<100*	**<200**	**<0.005%**

Please note: all figures are estimates based on the available data and should be treated as such. Minimum and maximum ranges have been included (in brackets) to account for margin of error. Figures marked with a * are either wholly or partly made from estimates of known types of organisations who may be in-scope as explained above the table, figures without a * come from analysis of the random sample of UK enterprises and extrapolating the figures. In assessing the potential impact of regulation on UK enterprises we are working on the assumption that an overestimate is more cautious, and therefore appropriate, than underestimating the impact.

Phase 4: How do organisations currently tackle online harm and how would regulation affect this?

Strategic sample and interviews with organisations

Phase 4 objectives

Understand the existing capabilities of in-scope organisations to tackle online harm and explore what the impact of potential regulation would be, through qualitative interviews with in-scope organisations.

The section covers:

- How organisations were identified for interviews
    - Defining the strategic sample
    - Recruitment and scheduling
    - Description of the final sample used for interviews
- Research objectives for qualitative interviews
- Summary of findings from qualitative interviews
- Detailed findings from qualitative interviews
- An estimation of the costs associated with regulation

Identifying organisations to interview

Defining the strategic sample

A strategic sample was used to select organisations to interview in this phase. Unlike the quantitative work in Phases 2 and 3, this is not a representative sample of organisations in the UK but instead a strategic sample of organisations who may be affected by regulation, and from whom there is most to learn in interviews.

Key considerations in developing the sample were:

- Prioritising organisations who present the **most risk** of potential harm (i.e. Tier 3 organisations). This was decided because these organisations are likely to be currently doing more to mitigate online harm (or one might at least expect them to be doing more than lower-risk organisations) and are likely to be most impacted by regulation.
    - It is worth noting that most of the Tier 3 organisations identified were *large* organisations, whereas only one *micro* Tier 3 organisation was identified. Therefore, targets for interviewing Tier 3 organisations were skewed towards medium and large organisations.
- Covering a range of other organisations in terms of **risk (i.e. tier)** and **business size** was also important given the OHWP's focus on **proportionately** (i.e. ensuring regulatory activities are proportionate to the risk and capability of organisations). This allowed researchers to speak to organisations with different capabilities and resources, as well as different potential for online harm, to explore how this affects approaches to preventing online harm.
    - It was decided to skew interviews towards those who might be expected to do more under proposed regulation, according to drafting documents from DCMS (see table below), as these organisations were more likely to have to make changes. Therefore, there were lower interview targets for Tier 1 organisations and micro and small Tier 2 organisations, given that

these types of organisation are unlikely to have to change their business practices in any significant way.

Table 13 showing potential requirements on different sizes and tiers of organisation under proposed regulation, according to drafted documents from DCMS at the time the research was carried out

		Organisation size			
		Micro (0 – 9)	Small (10 – 49)	Medium (50 – 249)	Large (250+)
Level of risk	Tier 1	• None	• Annual risk assessment	• Annual risk assessment	• Annual risk assessment
	Tier 2	• Annual risk assessment	• Annual risk assessment	• Annual risk assessment • Automated moderation tools	• Annual risk assessment • > 25 human moderators • Transparency reporting
	Tier 3	• Annual risk assessment • Automated moderation tools • Some human moderation	• Annual risk assessment • Automated moderation tools • Some human moderation	• Annual risk assessment • > 25 human moderators • Transparency reporting	• Annual risk assessment • > 50 human moderators • Transparency reporting

With these considerations in mind, targets were set to interview a total of 50 organisations across tier and business size:

Table 14 showing interview targets across tier and organisation size

	Micro	Small	Medium	Large	*Total*
Tier 1	1	2	2	2	**7**
Tier 2	2	2	7	9	**20**
Tier 3	1	3	9	10	**23**
Total	**4**	**7**	**18**	**21**	**50**

Other, secondary, sampling considerations included:

- **Type of platform:** ensuring a spread of different types of platform who fall into scope of regulation from different sectors, such as social media, online retail and gaming.
- **Type of business:** ensuring inclusion of different types of business, such as charities as well as limited companies.
- **Number of users:** ensuring that platforms with high numbers of users in the UK were prioritised in each tier, given that these organisations would have higher potential to affect a greater number of users. The most used platforms in each tier tended to correlate with the larger business sizes. We used data sources to provide information about the most used/visited sites in the UK[4].

[4] Amazon Alexa webpage & DomainTyper

- **Contribution to the economy:** ensuring we include the Tier 3 organisations who contribute the most to the economy—for example, those with large shares of the digital advertising market

Recruitment and scheduling interviews with organisations

Potential organisations to contact for interview were profiled into tiers using the scoring system established in Phase 3, and their business sizes checked using available data. This was often done by using the most recently published data on Companies House, information from their own websites, and other sources such as LinkedIn, Crunchbase and Owler.

Relevance to UK legislation was confirmed either through a Companies House listing or an advertised office in the United Kingdom. Because many of the organisations were multi-national, with those responsible for trust and safety or legal policy (i.e. those relevant to speak to for interviews) potentially located outside the UK, size refers to the global number of employees, with the exception of franchise-like organisations.

Organisations were contacted through a mixture of methods such as email, phone, contact forms on platforms' websites and direct LinkedIn messages. Organisations were provided with an information sheet regarding the aims of the project, what the interview would involve and how data would be used. In particular, all organisations were offered anonymity beyond being grouped by size, tier and, if appropriate, industry.

If an organisation agreed to take part in the research, an interview was scheduled either via teleconference or a face-to-face meeting was arranged.

Interviews lasted around 45-60 minutes, though some Tier 1 interviews were shorter. Interviewers followed the structure of a discussion guide, though each interview was adapted to be most relevant to the organisation in question.

Final sample of organisations who were interviewed

The table below shows the total interviews completed by tier and business size, as well as the total organisations contacted in each tier.

Table 15 showing the final sample of organisations who were interviewed across tier and organisation size

	Micro	Small	Medium	Large	*Total (contacted)*
Tier 1	1	2	0	1	**4 (29)**
Tier 2	1	1	3	5	**10 (63)**
Tier 3	1	1	3	11	**16 (26)**
Total (contacted)	**3 (20)**	**4 (22)**	**6 (29)**	**16 (47)**	**30 (118)**

As can be seen in the table above, a total of **118 organisations** were contacted for interview, and follow-up messages were sent if an organisation had not responded to initial contact. Overall, 25% (30) of the 118 organisations who were contacted agreed to and completed an interview.

The response rate from Tier 3 organisations was far higher than the response rate from Tier 1 or 2 organisations. This is to be expected given that regulation is likely to feel most relevant to these types of organisation, and thus worth their time to participate in the research. Additionally, large Tier 3 organisations were also most likely to have dedicated members of staff who work on 'Government Relations' or similar roles.

Regarding the secondary sampling criteria, the sample contained a mix of types of organisation, including:

- Social media
- Forums
- Review sites
- Blogging
- Gaming
- Retail

- P2P marketplaces
- Volunteering
- Official fans sites (e.g. official site for fans of a book)
- Job searching
- Fan fiction
- Search engines
- Accommodation searching
- Adult entertainment
- Dating

Similarly, although the majority of organisations who responded were limited companies, the sample also included a few charities and co-operatives.

Reliability of the qualitative sample

While the qualitative sample was chosen strategically and is not representative of all UK organisations, because of the small numbers of organisations in Tier 3 (the greatest potential for online harms as per current definitions, accounting for 0.005% of all UK enterprises), this strategic sample *does* in fact represent a relatively large proportion of these organisations.

If we repeated the exercise, we would need to speak to many (if not all) of the same organisations. For example, we interviewed 13 of the 16 most used social media sites in UK[5] who represent a large proportion of those likely to have to make changes under the OHWP. Similarly, if we weight by factors such as UK employees, share of the digital advertising market or declared profits, the proportion covered by the sample outweighs those not covered.

Qualitative research objectives

The main aim of the qualitative interviews was to establish, if possible, how much organisations may have to invest in order to meet the requirements of proposed regulation. This would enable an estimate of the potential cost to organisations in the UK as a result of potential OHWP regulation.

Specific objectives for interviews with organisations were to explore:

- The online activities they enable that carry risk of harm to users (as outlined in the OHWP – i.e. not cybersecurity or data protection issues)
- Their current practices and processes to mitigate that risk and to identify any harm occurring
- Their understanding of harms and how they may occur and an estimate of the level of harm they currently observe, either reported to them or proactively identified
- Where available, quantification of the associated **resources and costs** of practices and processes to identify and prevent harm (*Note: this is the primary research objective)*
- How these costs and resources would change if a duty of care was enforced

Limitations of the qualitative interviews

As we could not control which organisations responded to our request to speak with them, there is a bias towards those who are willing to engage. As previously discussed, we saw this in the fact that response rates are far better for Tier 3 organisations, who the research is highly relevant to, while Tier 1 and 2 organisations were less likely to reply. However, given that Tier 3 organisations will have the greatest requirements in the OHWP, this is unlikely to cause significant problems.

While those participating in the research were sent information about the topics to be discussed in advance of the interview, there was certain information that many were unable to provide, In particular, information relating to costs to mitigate online harm and the costs associated with potential regulation. In some cases those being interviewed simply did not have access to this data, and in others they were unable to share this

[5] Most popular social media platforms in the United Kingdom (UK) as of the third quarter 2021, by usage reach. Statista.

due to its sensitive nature. Many also felt they were unable to estimate specific costs associated with the proposed regulation in the OHWP without further clarification as to what this would entail.

Lastly, given that interviews relied on self-reported data, they are subject to the risk of bias.

Detailed findings

This section is split up into:

1. Approaches to identifying and dealing with online harms
2. The resources and costs associated with processes to identify and prevent harm, and how these may change under a duty of care
3. Attitudes towards the OHWP

1. Approaches to identifying and dealing with online harms

In general, the mitigations an organisation had in place were proportionate to the organisation's risk of potential online harm

As expected, different platforms experienced different types of online harm, and to varying extents. Almost all of the harms identified in the OHWP were brought up in discussions across interviews.

Generally, organisations running platforms where UGC and P2P were secondary to their function, such as reviews on a retail platform (i.e. Tier 1 organisations), were far less likely to have experienced any of the online harms in the OHWP compared to those whose primary purpose involved UGC and P2P (i.e. Tier 3 organisations).

Generally, the mitigations platforms had in place depended on the risk of potential online harm on that platform. Tier 3 organisations therefore had many more mitigations in place compared to Tier 1 organisations.

For example, human and automated moderation was present across all tiers, whereas processes such as reporting functions, paying for access to databases, such as Photo DNA, and publishing transparency reports, were only present in Tier 2 and 3 organisations.

The ways in which mitigations such as moderation were used varied across the tiers, with Tier 3 organisations more likely to use mitigations designed specifically for their platform

It is worth noting that different types of mitigation were implemented to varying degrees. For instance, while automated moderation was used throughout, the complexity and tailoring of this to the specific platform varied. For example, a Tier 1 organisation was using 'off the shelf' automated moderation to detect spam, whereas a Tier 3 organisation had developed their own bespoke automated software tailored to detect specific harms present on their site.

Organisations in Tier 2 and 3 had also tailored the options in their reporting functions to represent the harms commonly reported on their sites, and to enable them to better triage reports to ensure they dealt with the high priority harms first. For instance, reports of sexual harassment or CSEA would be flagged as high priority, while 'offensive' content or spam would be ranked as lower priority.

Some organisations mentioned the tension around putting specific options like 'sexual harassment' on their reporting options, given their concern that users might see this and assume it was something that took place on their site that they might encounter. However, it also helped the organisation to identify and deal with harms in a more efficient way.

The mitigations were also proportionate to the size of the organisation and quantity of UGC

As expected, there were differences in the capabilities of small versus large organisations—for example, one *large* job searching platform was using automated moderation to flag inappropriate terms, such as suggestions of nudity, self-harm and threatening language. Meanwhile a similar *micro* organisation did not have capability to invest in this.

However, there were other differences between these two organisations, which meant that automated moderation was not as necessary for the smaller platform. For example, the smaller organisation had under 5,000 users, while the larger organisation operates in 160 countries and has over 10 million users. For the latter, the quantity of UGC is vast, and automated moderation was required to keep on top of the content on their platform. Both organisations had a team of human moderators who would review any flagged content (either by users or automated software, or both).

Some were making decisions about whether the risk associated with certain features was worth the cost of mitigating against them

Some organisations were making decisions about whether or not the costs of moderating or mitigating harms associated with a certain feature was worth the benefits (i.e. increasing the number of users to their site or level of engagement with their site). For example, one Tier 1 small organisation who ran a fan site for a series of books they created used to host a forum for their 100,000 monthly website users. However, given the global fan base, and knowledge that children as young as 10 years old used their site, this required 24/7 moderation, which meant hiring an external moderation team. However, they decided this was costing more than the benefits the forum generated for the organisation, and removed this feature from their website.

Instead, they are now considering a feature in which users can upload drawings of characters from the books, with a caption. However, to enable all users to do this would, they estimated, cost around £20-30,000 in pre-moderation. Again, they did not think this was worth the investment, and so decided to limit the time period users can do this, reducing moderation costs to two to three days a month (approx. £2,500 a year).

Some felt there could be more guidance about how to deal with certain harms, particularly when the police need to be involved

A few organisations, such as a social media platform, a forum and a peer-to-peer marketplace, mentioned that they had struggled in the past with knowing how to involve the police, when appropriate (e.g. in situations where people appeared to be at immediate risk of harm).

When organisations had identified an illegal harm or a user who may be at immediate risk of harm, some struggled to know who to contact in the police, and how to contact them. Other challenges were when they had limited information about the user, given that people may have signed up with 'fake' details. This meant that all they were able to provide was an IP address. Some had 'worked out' that it was easiest to encourage the victim to report the crime and then provide the platform with their crime report number. They could then follow this with additional evidence, such as messages from their platform, sent directly to the police.

Others mentioned that any kind of 'standardised' advice on how to identify and deal with different kinds of online harm would be greatly appreciated. Some were looking to competitors for guidance on how to do this, while also using trusted resources from certain charities dedicated to specific topics, such as the Samaritans when it came to issues relating to mental health. This was particularly the case for the harms referred to as more 'subjective', such as hate speech, bullying and disinformation.

Almost all organisations were already motivated to invest in protecting their users for a variety of reasons

Most organisations were already investing in protecting their users in the absence of regulation and expected this investment would continue to increase over time. The reasons for investing in protecting users from online harm included:

- Creating a positive environment for users, to retain existing users and attract new ones
- To meet the requirements of advertisers and third-party suppliers, such as payment providers, who do not want to be associated with harmful platforms
- To remain competitive in the industry and keep up with their competitors

For the few who were not investing significant amounts in mitigation, some appeared to be ideologically opposed to the idea of collecting and moderating the data of their users. These organisations claimed to collect as little data as possible on their users for privacy reasons. This was referred to by more than one organisation as a US versus European perspective, with the US perceived to prioritise free speech and individual rights more than European countries.

These organisations struggled to estimate the cost of compliance with the OHWP, given that hiring additional moderators or developing/purchasing software would not in themselves result in compliance without a fundamental re-think of the service proposition. In some cases, this may mean their customers/users would leave.

2. The resources and costs associated with practices to identify and prevent online harm and how these may change under a duty of care

Discovering the costs and resources spent on protecting users from harm was a key area of discussion in interviews. Some organisations were reluctant to provide specific numbers or reported they did not have access to these.

It is important to note that the mitigations organisations had in place tended to reflect the level of online harm each organisation observed. Therefore, organisations dedicating lower levels of investment or resource to preventing online harm were not necessarily dealing with online harms any more or less effectively than those investing larger resources, as they might have had lower potential for online harm.

Each organisation did not provide the same details or use the same reference points when describing their costs, but the information given broadly fell into four primary areas:

- **Staffing levels**, including number of employees involved in safety functions and protecting users from online harm
- **Overall spending**—often a figure related to expenditure or revenue
- **Service-specific costs** spent on software or databases—for example, GIFCT and PhotoDNA
- Estimates for **spending to comply with regulation**, including to produce transparency reports

Staffing levels

Staffing levels refers to the number of employees whose work relates to content moderation and the protection of users, whether in part or exclusively.

The roles of staff involved varied depending on the size of the organisations. Large tier 3 organisations tended to have specific employees whose entire role was devoted to content moderation and user support in some form—with around 15-20% of their total workforce working in these teams. For smaller organisations, potentially those with less harm experienced, content moderation was one of a number of duties that a particular employee might have.

Many of these organisations used different titles for the staff responsible for protecting users from harm or moderating content posted on the platform. For anonymity, these teams are generically referred to as the organisation's 'User Safety' team, with more specific roles described if appropriate.

The table below sets out the findings by tier, the sector of the organisation's business and the organisation's size. Each bullet in 'Reported data' represents a single organisation.

Table 16 showing the number or proportion of employees whose work relates to content moderation or the protection of users, across the three tiers

Tier	Sector	Organisation size	Reported data
1	Publishing	Small	• An editor spent two-to-three days per month manually moderating content
	Retail	Small	• 20% employees worked on resolving 'customer issues'
	Reviews	Large	• 7.5% of workforce involved in content moderation

2	Marketplace	Small	• 16% of workforce closely involved in dealing with harms seen, with one employee working in this area full time, spending 20 hours dealing with reports and another five hours each week planning new functions and the next phase of automation
		Medium	• Around 30% of resource was dedicated to the team involved in supporting users with issues, which was the largest single team in the organisation and included the engineering teams involved in the development of AI
	Forum	Medium	• Just over 5% of its workforce devoted fulltime to moderation, supported by ten times that number of volunteer moderators • 15% of employees resourced to moderation, covering 7a.m. to 11p.m., with volunteer moderators covering hours outside this time
3	Social media	Micro	• Declined to provide exact numbers, but noted that 'one or two' people would be online at all times
		Small	• Every employee was involved in safety to some degree, but commented that there were never enough human moderators available
		Medium	•10% of employees in their user safety team, supported by a counselling service and relevant training
		Large	• 15% of employees working on content moderation was a figure given by two organisations, one of which qualified this by including engineers • Would not disclose the number of individuals, other than to comment that they had a 24/7 global user safety team
	Messaging	Medium	• Had a full-time user safety team comprising of just over 10% of total employees
	Gaming	Medium	• Equivalent to 50% of staff on user safety for one of its platforms, spending an hour each day dealing with user reports
		Large	• A team comprising 0.2-0.3% of global employees who deal with user reports
	Pornography	Large	• 3-4% of employees worked in compliance, which included data protection and security
	Forum	Large	• 10% of staff dealt with harms, including those in operations and engineering, with volunteers moderating certain parts of the site
	Online media and communications	Large	• Did not provide exact staffing levels, but explained that there were a variety of teams covering different services

Overall spending

For larger tier 2 and 3 organisations this refers to the amount spent on functions such as Trust and Safety or Community Experience, as well as other costs such as software or paying for access to databases such as PhotoDNA.

For smaller organisations, the costs become more specific and related to the time spent by individuals.

Table 17 showing the spend relating the protection of users, across the three tiers

Tier	Sector	Organisation size	Reported data
1	Training provider	Micro	• Cost of moderating comments was a 'negligible' part of business as usual
	Publishing	Small	• Had previously hosted forums costing a 'six-figure sum' for external moderation, but had deemed this too high
	Retail	Small	• £50, or half a day, per month on moderating reviews
2	Volunteering	Micro	• £2000 in total to deal with complaints, including costs of £250 for three days to develop a standard process for assessing reviews and £700 per year for an hour per week spent moderating reviews
	Forum	Medium	• 7% of annual expenditure on moderation
	Reviews	Large	• Annual costs of reported content moderation in the 'tens of millions of dollars'
	Dating	Large	• 'Hundreds of thousands' spent each month on content moderation and user safety
3	Messaging	Medium	• Budgeted 'in the low double digits' (i.e. approx. 10-15%) as a percentage of total company costs towards user safety in their 2020 budget
	Gaming	Large	• Content moderation as less than 1% of operating costs with the qualification that such estimations were difficult because content moderation formed parts of many people's jobs
	Social media	Large	• Equivalent to 14% of revenue spent on safety and security

Service-specific costs

These are the costs spent on particular services from other organisations, such as tools for content moderation.

Nine of the organisations interviewed referenced using the **IWF PhotoDNA** hash database to prevent child sexual abuse imagery, while seven reported using **GIFCT** to prevent terrorist content. These were predominantly **social media** organisations, though PhotoDNA was also used by the dating platform who were interviewed, for example. Varying costs were reported for gaining access to these databases:

Table 18 showing the spend on particular services from other organisations, across the three tiers

Tier	Sector	Organisation size	Reported data
1	Training provider	Micro	• Used free moderation software included alongside website hosting, and provided an annual £10 voluntary donation for the upkeep of the software
2	Forum	Large	• Estimated that tools for users to flag content would cost £12,000 per year
3	Social media	Small	• Had been quoted £10,000 for a PhotoDNA licence, and noted that there were many moderation companies offering software solutions at rates of $20,000 per month
		Medium	• Spent upwards of $50,000 per year on each of the PhotoDNA and GIFCT hash-sharing services

		Large	• Referenced a cost of $65,000 for GIFCT access
	Gaming	Medium	• Paid for Sift each month for both of its platforms, but did not provide a specific cost
	Pornography	Large	• Paid £70,000 to for the Internet Watch Foundation's services

Additional costs associated with regulation

These are the costs organisations estimated would be required in order to comply with potential regulation. Notably, the majority of organisations felt that the measures they had in place to address online harms were sufficient for regulation under their interpretation of the OHWP.

An exception was the introduction of **transparency reports**, though several large Tier 3 organisations, particularly in social media, already produced such reports. Other Tier 3 organisations, some smaller but also other large ones, did not produce transparency reports and stated that they did not have the data to do so. Some referenced the prohibitive cost of doing so, as well as the impact on public opinion and the reputation of their site should harms be made public.

Smaller and lower tier organisations did not produce transparency reports, but felt that the data could be gathered to do so. This would take time which could otherwise be spent on other work.

Table 19 showing the additional costs associated with regulation, across the three tiers

Tier	Sector	Organisation size	Reported data
1	Training provider	Micro	• Considered that compiling a report with numbers of reviews and moderated comments would be 'an afternoon's work'
2	Volunteering	Micro	• Described transparency reports as 'a pain' but suggested that these could be done for £50, or half a day, though this would be half a day that could not be spent on other work
	Marketplace	Small	• Already produced a comprehensive risk assessment and review every six months, taking several days and at a cost of £2,500-3000
	Forum	Medium	• Considered that current practices covered most of the suggestions in the OHWP, but did not object to committing further resources for specific tasks if this could be drawn from current moderation and development budgets
		Large	• Commented that producing a risk register or conducting an audit of harms would be feasible, though did not provide a cost
3	Social media	Small	• Stated that they lacked the infrastructure to produce a transparency report and added that an in-house lawyer would be necessary
		Medium	• Highlighted the costs of consultants and legal advice that would be required in response to legislation
		Large	• Two noted, without comment, that transparency reports were already produced • Another was producing a report added that regulation would not alter how much money was spent on protecting users from harm although meeting specific regulations would divert resources from other areas

			• Another was producing a report and commented that the cost of producing such a report was more of an issue than the content of a report
	Employment		• Produced an internal transparency report that was not made public

Miscellaneous costs

These are other costs that organisations referenced during interviews.

Table 20 showing other miscellaneous costs associated with regulation, across the three tiers

Tier	Sector	Organisation size	Reported data
2	Volunteering	Micro	• Had GDPR compliance incorporated into the build of their website, though preparing for compliance had otherwise taken 1-2 weeks of time
	Forum	Medium	• Had had a single user create thousands of accounts over nine years, costing them £13,000 in time spent banning these accounts, a further £5,000 on time spent engaging with the police, and an additional £2,000 in supporting other users who had been victims
3	Social media	Medium	• Noted that GDPR compliance had taken a year of their engineering team's time, with an associated opportunity cost
	Pornography	Large	• Estimated that competitor would have to spend £500,000 annually to reach their level of content moderation

3. Attitudes towards the OHWP

Most organisations were supportive of the need for regulation of some kind and had anticipated that regulation will happen at some point.

As expected, awareness and familiarity with the detail of the OHWP increased going up the tiers, with all of those in Tier 3 having processed the information in the OHWP. Although those in Tier 1 were often not aware of the OHWP, they generally understood the reasoning behind it and displayed a willingness to comply. However, they did not necessarily feel it was most relevant to them or their users.

There were some specific concerns and questions raised across the interviews. The types of concerns raised in Tier 2 and 3 interviews were:

- **Conflation of legal and illegal harms under a single system**: While most were comfortable with the duty of care covering illegal harms, some were concerned about 'conflating' these with the legal but harmful harms and felt that the duty of care would need to acknowledge that different approaches were required to deal with the two types of harms. Many wanted more clarity about legal harms and how to identify and deal with these. Bullying and disinformation were mentioned as examples of these types of more 'subjective' harms that were harder to deal with.
- **Ability to verify the age of users**: Many struggled to reliably detect which of their users were children, and indeed, most organisations claimed they would adopt age-gating without hesitation if a reliable service became available. One organisation with a user base including children and young people noted that they were more worried about compliance with the Age Appropriate Design Code as they felt this was more 'prescriptive'.
- **Balancing clear guidance with flexibility**: There was a tension between a desire for clear guidance, and having general principles to abide by. Knowing exactly what was required of them in particular circumstances would enable organisations to mitigate regulatory risk to themselves. At the same time, they also wanted the ability to implement measures they considered to be sufficient and appropriate to the risks on their platform, a 'principles' rather than a prescriptive approach.

- **Direct responsibility for content and legal consequences for individuals within a company**: Many mentioned their concerns about making directors liable for online harms on their platform. They felt 'liability' was different to a 'duty of care' and taking 'responsibility' for harms on their platform.
- **Time period for dealing with harmful content**: Some were concerned about implementing very strict timings in which organisations had to remove harmful content, such as within an hour. They felt this would cause organisations to withdraw from the UK.
- **Separate technology for UK operations**: Some who use the same technology across all the countries they operate in were concerned about having to make technological changes that applied only to the UK.
- **Reputational damage from transparency reporting**: Many already had the data and were doing risk assessments for internal use. However, some were concerned that if they provided the government with this data, the public would be able to request information via an FOI, which they worried could damage their reputation.

Note: Several organisations who were contacted after DCMS published the initial consultation response noted that they felt the government had listened to organisations. Others felt the government was engaging positively with the industry.

The types of concerns and questions raised in Tier 1—and some Tier 2—interviews were:

- **Redundancy of specific regulation in dealing with online harm**: Concern that it could become another bit of regulation they 'have to comply with', such as GDPR, but that the activities they would have to do wouldn't necessarily prevent their users from coming into online harm. This was often because they felt the likelihood of their users experiencing any of the harms on their platform was already low.
- **Compliance with the OH regulation taking resources away from running of their platform**: Concern that it would detract from other work they need to do. This especially raised by micro and small organisations who were concerned that spending additional time on compliance may take away from other work required to run their platforms. However, they did acknowledge that activities such as a risk assessment would not take up an excessive amount of time, and some reflected that this may be a useful activity for the business anyway.
- **Effect on organisations beyond social media**: One Tier 2 organisation mentioned that it felt like regulation 'designed for social media' and would negatively affect other types of organisation that it wasn't targeted at
- **Concern about the secondary impact on other platforms:** One micro Tier 1 organisation who used YouTube to market their services were more concerned about how regulation on YouTube would affect them in terms of additional work to ensure their videos were compliant with YouTube's rules—for example, manually marking whether each video was appropriate for children.

Estimating costs of regulation using Phase 2 and Phase 4 data

Method and sources

There were limitations as to how specific organisations can be about whether they will incur any additional costs, or what these costs might be, without knowing exactly what the regulation will require them to do. Therefore, any estimates of additional cost are based on current interpretations of the OHWP, and current practice within the organisations.

However, we were able to obtain three different types of cost estimate from organisations:

- How much an organisation that is already doing 'enough' (according to current interpretation of the OHWP) to mitigate harms currently spends on doing so, in terms of costs or resources.
- Specific costs associated with certain mitigation activities—for example, the cost of hiring X number of human moderators, or signing up to PhotoDNA. These can/have been calculated from desk research, and we were also able to collect these in interviews.

- How much an organisation not currently doing enough estimates it would cost them to comply. Without clear direction on what the regulation would require of organisations, most were unable and reluctant to provide or commit to any estimated costs. However, some were able to estimate these costs or estimate how much it would cost a competitor to put the same level of mitigations they had in place.

To reach a figure for the incremental cost of compliance (with potential new regulation), we must discount the expenditure on mitigation activities of two groups of enterprises:

- Organisations already compliant with what future potential regulation could look like/expect (i.e. doing 'enough' already according to current/reasonable interpretation of what the OHWP may entail)
- Organisations who would not be expected to take any further actions based on low risk of harms (i.e. Tier 1 organisations and smaller Tier 2 organisations), except potentially a risk assessment

Therefore, a rough estimate for the incremental cost to in-scope organisations can be calculated in the following way:

- Expected % of in-scope organisations requiring extra spend (i.e. those not excluded by current mitigation activities, size or low potential for harms), multiplied by the mid-point of estimates for incremental spend we identified in the research (we are using a figure of 7.5% of turnover[6])
  - For this rough calculation we have used an overestimate and assumed that *25%* of tier 3 organisations are required to take action, based on having spoken to a large proportion of tier 3 organisation in interviews
  - Similarly, based on conversations with tier 2 organisations we have overestimated that 10% of organisations may be required to take actions, beyond producing a risk assessment
- Plus, a generic cost for risk assessment for every in-scope organisation if this is a requirement for all in-scope enterprises

Overall cost to organisations who fall in scope

Taking this approach, we can use turnover estimates from BEIS[7], our estimates for the number of in-scope organisations within each tier and business size category, and estimates from the qualitative interviews to determine how many of each might have to invest more (i.e. are not already doing enough). This provides a broad range of approximately <£1 billion - £2 billion in total. The largest proportion of this estimate comes from the assumption that many medium and large Tier 2 organisations will be expected to bring in the highest standards of harms mitigation/moderation.

The table below gives the number of organisations in each tier and of each size who are expected to fall in scope, alongside cost estimates for compliance with regulation for each tier. Green shading for Tier 3 organisations and medium and large Tier 2 organisations indicates the groups of organisations expected to have to make changes. Not every organisation within these groups would have to make changes, as this would depend on their current levels of practice.

Please note these estimates were provided in the context of businesses' interpretation of the OHWP at the time the research was conducted, i.e. the cost of additional content moderation for businesses required to address all harms in the OHWP including extra protections for children. As outlined in the draft Online Safety Bill, duties related to legal but harmful content and activity as accessed by adults will only apply to the largest and highest risk services (Category 1 platforms). These costs are therefore likely to overestimate the incremental cost of regulation; however, they are still useful to provide an indication of the likely scale of impact.

[6] The range of estimates we received from the organisations we engaged with were between 1 and 14-15% of revenue. As we do not have figures for revenue of UK organisations, we are using turnover as a proxy.

[7] Business Population Estimates for the UK and Regions 2019.

Table 21 showing estimated overall cost to organisations who fall in scope, across the three tiers, including exemptions

Original figures, including Autumn 2020 exemptions							
	Micro	**Small**	**Medium**	**Large**	***Total***	**Potential action**	**Cost**
Tier 1	8,500	500	1,000	400	~10k	Tier 1 organisation not required to take further action	£0
Tier 2	8,500	500	1,000	100	~10k	Approx. 10% of larger Tier 2 organisations required to take actions	~£280[8] million
Tier 3	-	-	400	50	~10k	Approx. 25% of Tier 3 organisations required to take actions	~£320[9] million
Total estimated number of in scope organisations	17,000	<3,000	<1,000	<1,000			~£1 billion
Average turnover (mean)	£0.16 million	£2.85 million	£16.16 million	£211.8 million			
Please note: These estimates are based on a number of different estimated figures and should be treated with caution. As noted previously, the actual cost to organisations of potential regulation depends on a wide range of factors, not all of which can be accounted for in the figures presented here. These should be used as a guide only.							

Example calculation using medium Tier 2 orgs: 10% of 1,000 orgs (100) x 7.5% of avg. turnover for medium size enterprises (0.075 x £16.16m) = £121.2m

Table 22 showing estimated overall cost to organisations who fall in scope, across the three tiers, pre-exemptions

Original figures, pre-exemptions							
	Micro	**Small**	**Medium**	**Large**	***Total***	**Potential action**	**Cost**
Tier 1	c.120,000	c.16,000	c.6,500	c.1,000	c.143,500	Tier 1 organisation not required to take further action	£0
Tier 2	c.28,000	c.5,000	c.3,000	c.1,000	c.37,000	Approx. 10% of larger Tier 2 organisations required to take actions	c.£1-6 billion
Tier 3	c.<10	c.<10	c.<50	c.<100	c.<200	Approx. 25% of Tier 3 organisations required to take actions	c.£825 million
Total estimated number of in scope organisations	**150,000** (20 – 270k)	**25,000** (8 – 35k)	**<10,000** (1.5 – 15k)	**<3,000** (<1 – 3k)			c.£1-7 billion
Average turnover (mean)	£0.16 million	£2.85 million	£16.16 million	£211.8 million			
Please note: These estimates are based on a number of different estimated figures and should be treated with caution. As noted previously, the actual cost to organisations of potential regulation depends on a wide range of factors, not all of which can be accounted for in the figures presented here. These should be used as a guide only.							

It is very important to note that the actual cost of any potential regulation depends on many factors, which the above estimates do not account for, and there are a number of reasons the above estimates may be considered cautious overestimates. For instance:

- It is based on the **highest known costs** from our qualitative interviews
- Estimates for the **number of in-scope organisations**, as noted previously, tend towards being **overestimates**
- It assumes that organisations are either doing enough or not, and for those who are not the cost would be the same, when **in reality** the **distinction is less binary**

[8] Range between 1% incremental cost and 15% incremental cost = £37 million to £560 million

[9] Range between 1% incremental cost and 15% incremental cost = £43 million to £640 million

- If the cost were this significant, **some organisations may choose to change/reduce their features**/functionality
- The **cost of mitigation** processes/practice may **reduce over time** with off-the-shelf moderation products, controls built into platforms, or improvements in internal capacity/efficiency
- The cost assumes these measures would be taken for compliance purposes only, whereas we know that many organisations employ moderation and **harms mitigation practices for wider commercial reasons**

<£1 billion - £2 billion provides a very broad guide for what the potential cost of regulation to be, but it is important to highlight this range as there are still many unknowns, as detailed above. On the whole, we expect the real cost accrued by UK enterprises due to some form of new regulation would be relatively small, restricted largely to the small number of organisations who have the highest potential risk based on the features of their sites/apps—many of whom are already putting in place reasonable measures to mitigate online harms.