

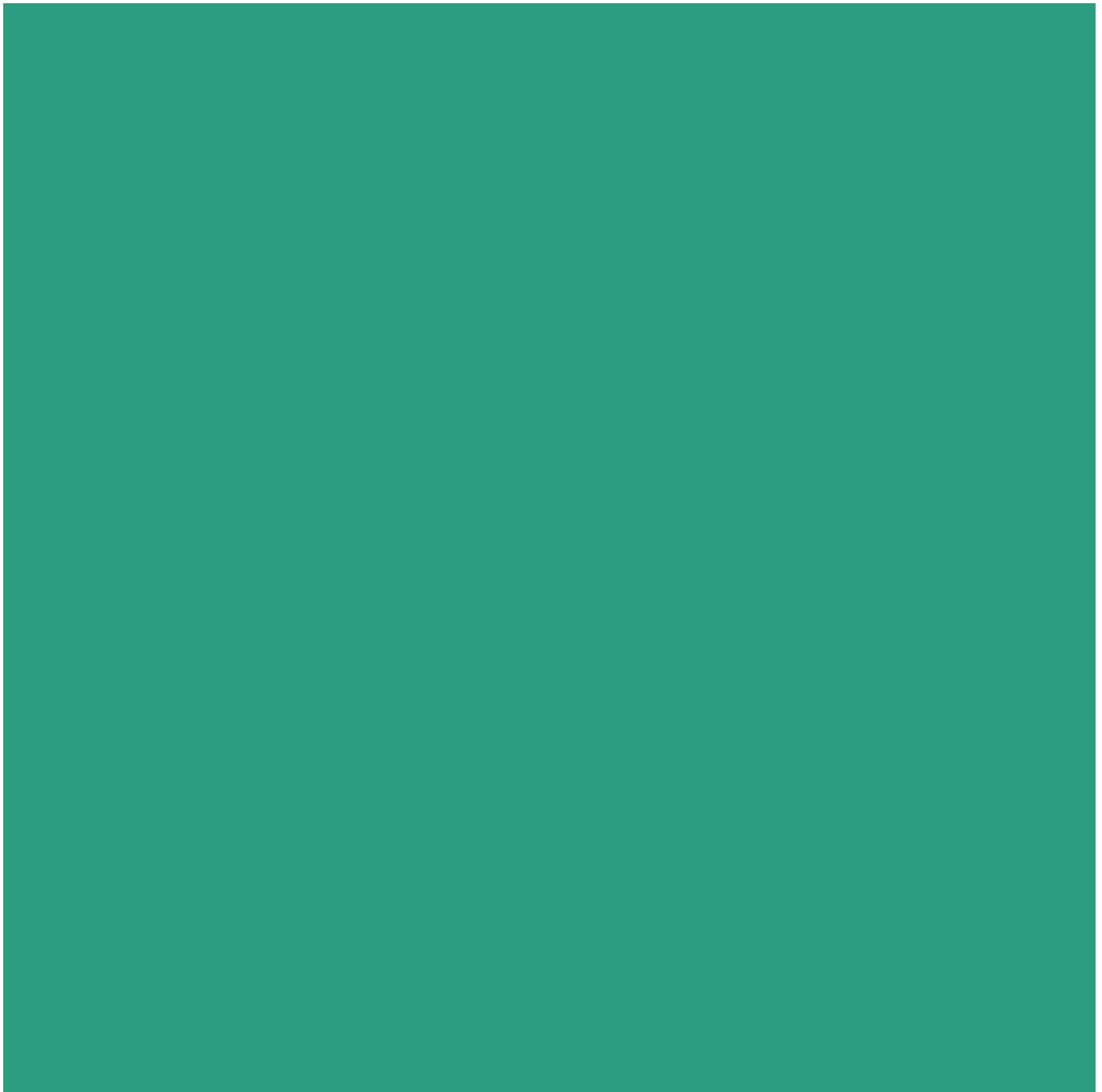


Department for  
Digital, Culture,  
Media & Sport

REVEALING REALITY

August 2021

# Online Harms Feasibility Study



# Preface

This research was commissioned by DCMS to explore the feasibility of measuring the prevalence and impact of online harms.

It aimed specifically to explore the measurement of four areas of interest, where there is currently no established approach to measuring harm. These harm areas were selected as they represent a diverse group of harms, ensuring the research was able to explore as many approaches to measurement as possible.

These selected harm areas were in the 'legal but harmful' space:

- Cyberbullying
- Online abuse
- Inappropriate access to content
- Promotion of illegal / risky / dangerous behaviour, notably the promotion of eating disorders, self-harm, and suicide.

## Method

This research explored what currently exists to measure the prevalence and impact of each harm, and where improvements could be made to enable accurate online harm measurement. It included:

### Consulting experts:

- Qualitative interviews were carried out with **24 experts** including **academics, charities, online platforms and Safety Tech firms**<sup>1</sup>.
- Interviews with experts explored the different ways that the selected harms are currently measured, the key challenges in doing so, and what could be improved.
- Follow up interviews were carried out with seven experts, who provided feedback on the initial research findings.

### Mapping existing measures:

- This research reviewed existing measures for the four harm areas.
- Measures were identified through reviewing academic literature, reports from the government and charity sector, and platform transparency reports.
- The relative advantages and disadvantages of each measure were evaluated against:
  - A **framework** breaking down the **different components of 'online harm'** developed during the project. It is outlined [here](#).
  - **Validity**: How accurately it measures what it is intended to measure?
  - **Pragmatism**: How easy would it be to use this measure?

While speaking to experts and reviewing the existing measures for the four harm areas, some overarching challenges, concepts and suggestions for how to measure online harm were identified. These are summarised in **Section 1** of the report. This section has implications for online harm measurement beyond the four harms areas specifically explored during this research.

## The report contains 3 sections:

### Section 1

summarises the key findings and recommendations for measuring online harm. It introduces some fundamental concepts and models which are referred to throughout the report.

### Section 2

provides detail on the methods currently used to measure online harm, and their relative strengths and weaknesses.

### Section 3

provides detail on the measures which currently exist in each of the four harm areas, as well as recommended next steps for improving them.

---

<sup>1</sup> GOV.UK. 2021. [The UK Safety Tech Sector: 2021 Analysis](#). [online]

# CONTENTS

<b>Section 1:</b>		<b>5</b>
Key findings and next steps for measuring online harms	Why measurement matters	6
	A model for conceptualising 'online harm'	7
	What currently exists?	11
	What could be?	18
<b>Section 2:</b>		<b>25</b>
Methods currently used to measure online harm	Transparency reports	26
	Surveys	30
	Qualitative research	35
	Automated tools (Artificial Intelligence, machine learning)	39
	Public / government data sets	43
	Public reporting sites and helplines	45
<b>Section 3:</b>		<b>47</b>
Measurement in the four harm areas	A. Online abuse	49
	B. Access to inappropriate content (specifically pornography)	55
	C. Cyberbullying	60
	D. The promotion of illegal / risky / dangerous behaviour	66
<b>Appendix 1:</b>		
	List of sources reviewed when mapping the measures in each harm area	75

SECTION 1:

# **Key findings and next steps for measuring online harms**

This section summarises what currently exists to measure online harm, the key challenges in doing so, and what the future of measurement should involve.

It introduces a model for conceptualising 'online harm' and a process for effective online harms measurement, which is referred to throughout the report.

## Why measurement matters

The draft Online Harms Safety Bill<sup>2</sup> is a pioneering step towards ensuring that tech companies better protect people online, whilst defending individual freedom to explore and express oneself.

As with any new regulation, the crucial question will be: Is it achieving its goal? Is the amount of harm that people come to online reducing, and are people still able to experience the many positives of being online while being better protected from the negatives?

As the appointed regulator, Ofcom will need to be able to independently monitor how well current approaches to mitigating online harms are working, and hold the relevant parties to account for what happens as a result of their products.

How well Ofcom can do this depends significantly on how well 'online harm' can be measured and tracked. Getting measures wrong could have unforeseen consequences that limit the ability of interventions to have their intended impact on reducing harm—e.g., focussing on the wrong areas, over or underestimating the relative impact of any given intervention or obscuring what is actually happening.

Until now, no one has set out to establish a comprehensive way to track the total amount of online harm over time. Government, regulators and charities have relied on piecemeal indicators of different types of harm—from transparency reports to one-off studies on topics such as cyberbullying, none of which has been set up with the explicit intention of measuring harm over time.

With the development of Online harms regulation and the appointment of a regulator with powers and responsibilities to track online harm over time, there is a critical opportunity—arguably a duty—to get the measurement of online harms (and benefits) right.

## What do we mean by 'online harm'?

An 'online harm' is just a harm which can be attributed to something that happened online.

Some academics would argue that "there is no such thing as an online harm" as the online world is not separate from 'real life'. Sometimes the cause of a harm sits in the online world, and sometimes it sits in the offline world, but the experience itself of being harmed is not online.

If a child has low self-esteem as a result of bullying online, while the cause is online, the impact, or harm, is not.

To measure 'online harms', both the **real-world impact** and the **cause of the harm** originating in the online world need to be taken into account.

When reviewing the range of research and efforts to measure 'online harms' it was apparent that for the most part what is being measured are things that have the potential to cause harm or can be assumed to cause harm in the online world, such as 'abuse', or 'content relating to self-harm'. But often whether they have definitely caused harm or how much harm they have caused is unknown. Harms vary in severity, in who they're harming, and in what contexts.

Removing these assumed or potential causes of harm is likely to be part of the solution to reducing harm overall. However, measuring assumed causes of harm without strong evidence linking it to harmful outcomes—and the context in which harmful outcomes occur—does not provide a reliable or accurate measure of actual harm.

It is important to remember that the same online experience can have both negative and positive outcomes—for different people, at different times or in different contexts. For example, content relating to self-harm may

---

<sup>2</sup> Department for Digital, Culture, Media and Sport, 2021. [Draft Online Safety Bill. Department for Digital, Culture, Media and Sport.](#) [online]

make some individuals feel supported or be educational in raising awareness about the issue. It may also be distressing or encourage negative behaviours for others.

Without accounting for the positive outcomes, as well as the negatives, there is a risk that actions or content perceived as harmful may be responded to disproportionately and indiscriminately. For example, if all content referencing 'self-harm' is removed, all the positive discussion and support that can be helpful to some people—and to society's understanding of the issue—would disappear too.

## A model for conceptualising 'online harm'

### *What we can learn from Health and Safety*

To measure how much harm is being experienced, the nuances of what actually constitutes 'online harm' needs to be better articulated.

The Health and Safety sector provides a useful parallel for this. Just like 'online harms' prevention, the role of Health and Safety is to protect people from harm.

This research has borrowed the Health and Safety concept of **Hazards, Risks and Harms**, and used it as a framework for breaking down 'online harm'. Indeed, the language of 'Hazards' and 'Risks' has been used before in the online harms world<sup>3,4</sup>, where academics have tried to break down what is meant by 'online harm'.

So, what is meant by Hazard, Risk and Harm?

Take the example of petrol—a **hazardous** substance, whose qualities mean that it has the potential to cause harm. However, it won't necessarily do so. It depends on **risk** factors: Where it is, how it is stored, what it is being used for, and by whom. Depending on these risk factors, it has the potential to do great harm—start a fire, injure or even kill people.

But its negative impacts can also be minimal.<sup>5</sup> There can even positive ones such as enabling people to get around in their cars.

If the prevalence of petrol in the world was measured, it would tell us almost nothing about the 'petrol related harms'. In the same way, measuring the prevalence of certain types of online content tells you very little about the level of harm being caused.

It is worth noting that where there is clear evidence linking online content with harmful outcomes, it becomes more feasible to use measures of online content as a proxy for harm. Indeed, some hazards will be more closely linked to harm than others.



**HAZARDS**  
STIMULUS



**RISK FACTORS**  
CONTEXT




**HARMS**  
IMPACT

<sup>3</sup> Livingstone, S., 2013. [Online risk, harm and vulnerability: reflections on the evidence base for child Internet safety policy](#). ZER: Journal of Communication Studies, 18 (35). pp.13-28. [online]



<sup>4</sup> Vidgen, B., Burden, E. and Margetts, H., 2021. [Understanding online hate: VSP Regulation and the broader context](#). [The Alan Turing Institute](#). [online]

<sup>5</sup> Note: aside from negative environmental impacts

Applying this to online harms, the table below illustrates what we mean by 'hazards', 'risks' and 'harms'.

Concept in the Hazard, Risk, Harm model	Description	Example relating to the topic of 'the promotion of eating disorder content'
 <p><b>HAZARDS</b></p> <p><b>Online experiences that are a potential source of/route to harm.</b></p> <p>Hazards do not always cause harm – it depends on the risk factors.</p>	<p><b>There are different types of hazard:</b></p> <ul style="list-style-type: none"> <li>■ <b>Content</b> that someone is exposed to E.g. pornographic images</li> <li>■ <b>Interactions</b> (with another user) E.g. bullying behaviour</li> <li>■ <b>Design features</b> E.g. constantly refreshing feeds, ability to connect with strangers, 'like' buttons For example, the sheer amount of time people spend online is encouraged by design features such as the constantly refreshing feed, and contact with strangers can be facilitated by the fact that a platform is designed in such a way that allows children to connect with unknown adults.</li> </ul> <p><b>Some hazards will be inherently more likely to cause harm than others:</b></p> <ul style="list-style-type: none"> <li>■ <b>Legal vs illegal hazards.</b> Many serious hazards have already been rightly classified as illegal – these hazards (e.g. terrorist propaganda, child sexual abuse imagery) have a much greater recognised risk than those classified as 'legal but harmful', and are dealt with differently</li> <li>■ Within the 'legal but harmful' hazards—such as content promoting eating disorders—there will likely be some types of content that are more or less likely to cause harm</li> </ul>	<p>May vary from content displaying 'diet tips' to content explicitly encouraging anorexia</p>

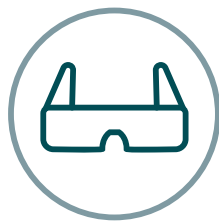


Concept in the Hazard, Risk, Harm model	Description	Example relating to the topic of 'the promotion of eating disorder content'
 <p><b>RISK FACTORS</b> Things that change the likelihood that a hazard will cause harm to an individual</p>	<p><b>Risk factors are based on who is exposed to the hazard and in what context</b></p> <ul style="list-style-type: none"> <li>■ The <b>situation</b> or context the hazard occurs or exists in (where, when, how)</li> <li>■ The <b>individual</b> / 'victim' (two people may encounter the same hazard, but only one experiences harm)</li> <li>■ The <b>level of exposure</b> (how much and how often the person is exposed to a hazard)</li> </ul>	<p>An individual's previous experience of eating disorders, age and gender</p>
 <p><b>HARM</b> The negative consequence on someone resulting from a hazard combined with risk factors</p>	<p>There are many <b>types of harm</b> that could occur.</p> <ul style="list-style-type: none"> <li>■ <b>Internalised harm:</b> Changing how you think and feel.                             <ul style="list-style-type: none"> <li>□ This may be immediate (e.g. being upset) or a less recognisable change in how you think, which could have negative future consequences (e.g. a changing perception of body image)</li> </ul> </li> <li>■ <b>Externalised harm:</b> Changing what you do.                             <ul style="list-style-type: none"> <li>□ Behaviours which relate directly to the hazard (e.g. self-harming), or behaviour that has a negative knock-on effect (e.g. retreating from social life / deleting social media)</li> </ul> </li> <li>■ <b>Societal harm:</b> The impact a hazard could have on society.                             <ul style="list-style-type: none"> <li>□ Such as changing perceptions or attitudes that can become unhealthy or lead to negative outcomes for others. This may include the normalisation or desensitization to certain content or behaviours</li> </ul> </li> </ul> <p>Harms also <b>vary in their severity</b>. For example, from causing mild, temporary distress to contributing to someone self-harming.</p>	<p>Emotional distress, triggering existing issues, encouraging damaging body image views</p>

## Harm can manifest immediately, but also accumulate over time

While some harms may affect someone immediately, others may manifest in the long term. For example, a child being contacted by a dangerous adult online would constitute an immediate risk. One instance is enough to cause harm, immediately. However, the long-term exposure to unhealthy body ideals is a health issue. One exposure might not do much, but over time, being immersed in it could build up to cause serious long-term harm. Similarly, design features which encourage users to spend increasing amounts on a given platform may not lead to any instant negative consequences but could contribute to sleep problems and poorer mental health over a period of time.

Again, this has parallels with the Health & Safety model, which includes both **safety** issues (immediate, obvious harm) and **health issues** (longer term, accumulative harm).



**SAFETY ISSUES**



**HEALTH ISSUES**

### Summary

- An 'online harm' is just a harm which can be attributed to an online experience (whether this relates to content, behaviours or other activities).
- It is a broad and complex issue which leads to significant measurement challenges.
- To provide greater clarity for measurement and analysis, we recommend conceptualising the different components of online harm as 'hazards', 'risk factors' and 'harms' (borrowing from the Health & Safety sector).
- Therefore, measures must account for hazards, risks and harms. They must be clear about which one they are measuring and the limits of what that information can demonstrate (e.g. knowing the prevalence of hazards does not necessarily indicate the prevalence of harm).
- Measurement of online harm must account for both positive and negative outcomes as a result of online experiences. It must not ignore the positive impact that certain hazards can have, and where the same hazard may have a different impact depending on who it is exposed to, and in what context.
- Measures of hazards must focus not only on content, but also on online behaviours and the design of services and platforms that determine what people do and how they do it.
- Measures of harm must focus not only on harm to the user, but also on wider societal harm: Individual and collective harm.

## What currently exists

### *The current landscape of 'online harms' measurement*

A variety of methods have been used to indicate the **prevalence** and **impact** of different 'online harms'.

**Prevalence:** How many times a hazard or harm has occurred

- E.g. how many times have people been exposed to content that promotes self-harm (a hazard)?
- E.g. how many times have people felt distressed (a harm) by seeing content that promotes self-harm?

**Impact:** To what degree harm (or a positive outcome) has been experienced

- E.g. knowing the level of distress caused by someone being exposed to certain types of content
- E.g. understanding the degree to which online experiences influence someone's offline behaviour

These methods include transparency reports published by social media organisations, quantitative and qualitative research into people's online experiences, and studies utilising automated tools to analyse social media and other platforms to identify different 'harmful' content. Research and measures have been developed by a range of stakeholders in the online harms space, by the platforms themselves, academia, charities, regulators, and government.

Broadly, most measures fall into these categories:

Method	Description	Example
<b>Platform transparency reports</b>	Reports published by social media platforms containing information about the platforms' guidelines and policies, and data on how these are enforced.	Five million pieces of content relating to Suicide and Self-Injury were actioned in Q1 2021. <a href="#">Facebook transparency report</a>
<b>Surveys</b>	Surveys of the general public, and of specific groups who may be more likely to experience certain 'online harms'. Children, for example.  These tend to be conducted by charities, academics, and regulators (e.g. the ICO and OFCOM).	27% of young people reported being 'cyberbullied' in the previous 12 months.  <a href="#">Ditch the Label Annual Bullying Survey (2020)</a>
<b>Qualitative research</b>	Data collected first-hand from researchers in interviews, focus groups, and digital ethnography (observing what people are doing online).  This type of research tends to be conducted by charities, academics, and regulators.	Some young people felt online pornography had led to the copying of "rough" sexual behaviour.  <a href="#">BBFC, Young People Pornography &amp; Age Verification (2020)</a>

Method	Description	Example
<b>Automated tools</b>	<p>Using automated tools (AI, machine learning) to identify content or behaviour.</p> <p>These tend to be used by academics, platforms and Safety Tech firms.</p> <p>Academics often use automated tools when analysing samples of social media data (e.g. a dataset of tweet), to detect the prevalence of potentially harmful content. These are referred to as 'measurement studies'.</p>	<p>In June 2020, 4.4% of all replies to MPs' tweet were 'abusive'.</p> <p><a href="#">MP Twitter Engagement and Abuse Post-first COVID-19 Lockdown in the UK</a></p>
<b>Official government / public sector data sets</b>	<p>Data collected by government departments or public sector organisations (e.g. the Police or NHS).</p>	<p>According to data reported in 2017/18, from 30 out of 44 police forces, 1,605 online hate crimes were recorded in England and Wales. This accounts for around 2% of all hate crimes.</p> <p><a href="#">Hate Crime, England and Wales 2017/18</a></p>
<b>Public reporting sites &amp; public helplines</b>	<p>Public websites or helplines publishing data on the concerns / reports they receive in their interactions with members of the public.</p> <p>Helplines and sites are often run by charities.</p>	<p>Self-harm was one of the top three topics on the Childline message boards.</p> <p><a href="#">Childline annual review 2018/19</a></p>

## Current measures are extremely limited and present several key challenges

While a variety of measures and sources exist, they do not provide an accurate or consistent enough picture of online harm to reliably measure the prevalence or impact over time—there is no 'ready-made' solution.

There are numerous challenges to measuring online harm well. Collecting the breadth and depth of data required to make accurate estimates of the volume and impact of online harm is a significant undertaking.

Existing challenges and barriers identified in this work are important to highlight as they indicate key 'watch-outs' and requirements for best practice by mitigating these issues. Broadly, these challenges / barriers fall into several categories:

- #1. Most potential measures / sources were not designed with the intention of being used as a measure of online harms
- #2. Technical limitations of research methods employed
- #3. Lack of consistent definitions and granularity

## #1. Potential measures not intended for use as an online harms measure

### Most 'measures' have not been established primarily as a way to measure online harm

Few of the measures identified in this research have been established specifically to track online harm over time.

The range of existing measures—from social media platforms' transparency reports to one-off studies on topics such as cyberbullying—provide a wide range of potential indicators of different types of harm or aspects of online experiences. But none have been set up with the explicit intention of measuring the total amount of online harm. They have been developed for a wide range of reasons, and their use as a measure of online hazards or harm is largely secondary—a potentially useful proxy, rather than a dedicated measure.

For example, transparency reports provide information about the action taken by platforms on content which violates their community guidelines. The development of these guidelines and reports were likely driven by a range of factors including protecting users, publicity, and politically or socially relevant issues based on the locations they operate in. While the data these reports provide may be a useful indicator about some harm areas, its primary purpose is not as an entirely impartial measure of harm. Although, as covered in a later section, there are opportunities for transparency reports and the data that sits behind them to provide an important contribution to understanding online harm.

*"It's all driven by our community guidelines, so all our efforts go into seeing if something does or doesn't contravene our guidelines"*

Social media platform

Other measures such as public helplines are set up primarily to support people, so understandably are unlikely to prioritise or be able to track the interactions they have relating to 'online' harms.

*"We do get insights about what is going on from the calls, the volunteers feedback about them, but it's anecdotal"*

Helpline

### A lack of ongoing measurement

Many studies (e.g. surveys and qualitative research projects) are one-offs, not repeated and sometimes not designed with tracking in mind. They are unable to observe change over time.

## #2. Technical limitations of the methods

The research methods and tools available have numerous limitations, making the use of multiple methods a necessity in order to understand the different aspects of online harms appropriately. Section 2 contains a more detailed review of key methods; the pros and cons; and examples from different online harms areas, but some overarching challenges have been described here.

### Many measures rely on self-report data from internet use

Measures which rely on self-reported data—someone telling you what they've experienced online and how it has impacted them—have several limitations.

### Low recall accuracy and subjective interpretation of questions

People struggle to accurately recall what they have been exposed to online, what they have done online, and even how they felt in the past. If someone is asked whether they have “come across ‘trolling’ in the last year”, not only would it be hard to recall all the content they have seen in the last year, but it is likely that each person will have a different interpretation of what constitutes ‘trolling’. People’s interpretations of questions and concepts are subjective.

### Social desirability bias

People may also feel pressured to answer in ‘socially desirable’ ways. For example, a child completing a survey about whether they have seen pornography on social media may choose to lie, particularly if their parent is completing the survey with them, as is often the case.

### Unable to report on harms that a user does not recognise

Self-report cannot be used to gauge harms that people don’t recognise have happened to them. For example, while a young person watching pornography may feel that it is not harming them in any way, it may be negatively shaping their perception of sexual behaviour, unbeknownst to them.

*“People don’t always know whether something is having a negative impact on them, at the time it may seem positive, but in the future, they might reflect that it was actually bad for them”*

Helpline

### Measures tend to focus on short-term impacts

It is easier to ask people about short-term impacts like ‘how upset did this content make you feel?’ than long-term impacts that may accumulate or manifest over time, and are therefore less easy to attribute to specific online experiences. For example, changes in perceptions of body image, and perceptions of normal sexual behaviours are likely to manifest over time, and it will be difficult for an individual to identify how their online activity has contributed.

While qualitative measures are better able to explore more nuanced, long-term harms, to establish causal links between online activity and longer-term outcomes, longitudinal studies—in which the same person is tracked over a period time—would be needed. Currently, there are few longitudinal studies in the harm areas explored.

### Automated tools are not able to understand context

Automated tools in this area (using techniques like machine learning and functions that would be described as artificial intelligence) are predominantly limited to identifying certain types of online content (hazards). For example, key words known to be associated with abusive behaviour like racial slurs and swear words, or images and videos that contain graphic content. These tools do not have the ability to identify and interpret context (risk factors) effectively, or to identify where harm has actually occurred. While their ability to process large quantities of data will be invaluable, they must be deployed as part of a wider programme.

*“If it doesn’t contain an abusive word, it [AI] won’t pick it up, so understanding context is difficult”*

Academic researching online abuse

For example, automated tools are not able to detect where harm has occurred, and to what degree. While an AI may identify that 10% of tweet to MPs are classified as 'abusive', other methods are required to understand the impact of these tweet—what type of content has been particularly harmful, how severe has the harm been, and how are their friends/family and others who have seen the tweet online affected?

*"We can't be sure that the MP has seen the tweet or how severely they have been affected"*

**Academic researching online abuse**

These challenges are well documented. The Alan Turing Institute's report on measurement of Online Hate<sup>4</sup> highlights that automated tools or AI are not a 'silver bullet' for measuring online harm. This report lists challenges in using automated tools, in relation to online hate specifically, but are also relevant for other areas of online harm. They include - AI lacking understanding of the wider social and historical context; lacking understanding of the speaker's identity; performing poorly on video, images, memes and audio compared to text; and being difficult to update over time as expressions of online hate changes.

Platforms also raised the challenges they face in relying on AI tools to detect hazards, and the need for user reporting to identify more nuanced instances of their community guidelines being broken.

*"User reporting and moderation picks up things AI never could, sometimes what people are doing doesn't register as abusive with AI, but the context in which it's being done is. People are creative in the ways they can be horrible online"*

**Social media platform**

### #3. Lack of consistent definitions and granularity

#### Inconsistent definitions of online harms used in different measures

Attempts to measure the same concept—e.g. the prevalence of 'cyberbullying'—often use different definitions. For example, some surveys will give detailed explanations of what they would consider to be 'cyberbullying' while others ask respondents to use their own definition of the term. It has been recognised that there are often no international definitions of concepts relating to online harms<sup>6</sup>. This has significant implications for accurate measurement and attempts to compare and contrast data. Different platforms also use different definitions for similar areas of online harms, writing their own community guidelines rather than having an official definition to work from to ensure consistency across platforms.

*"You can't say transparency reports are comparable across platforms, if you looked at cyberbullying and it had gone up or down in our report, that could just be a result of us getting better at finding it"*

**Social media platform**

---

<sup>6</sup> United Nations, n.d. [The United Nations Strategy and Plan of Action on Hate Speech](#). United Nations. [online]

### A lack of granularity or detail about what hazards people are exposed to

Measures often provide limited detail in describing what exactly people have been exposed to online. For example, platforms and surveys might report how much ‘online abuse’, ‘self-harm/suicide content’ or ‘pornography’ they have removed, or people report seeing. However, there is a large spectrum of content of varying degrees of severity or concern that could fall into each of these categories—with some likely to be far more closely associated with harm than others.

*“We report how quickly we get to the content, how much of it there is, but not who is seeing it or how long”*

**Social media platform**

This is particularly the case with areas of online harms such as underage viewing of pornography. Because this behaviour is illegal, it is assumed that all underage viewing of pornography causes harm, and there is no differentiation between different types of pornography. This may hide the fact that certain types of pornography are likely to be more closely linked with harmful outcomes than others. Hypothetically, knowing that 20% of those viewing pornography underage experience harm, and the overwhelming majority of these people are seeing a certain type of pornography, has far greater implications for targeted and effective harm prevention than simply knowing that 20% experience harm.

### Few measures account for how often people are exposed to hazards

Similarly, few measures account for the frequency of being exposed to a hazard, and how this interacts—with the severity of harm. For example, it could be the case that seeing extreme content once is less harmful overall than being frequently exposed to less severe content. Capturing the patterns of exposure or behaviour / experiences of users online is just as important as capturing what kind of things people are exposed to.

*“The serious pro-ana content is seen by a small number of people, the endless bikini pics online might be having a bigger impact on more people over time”*

**Academic researching eating disorders**

### There is limited detail around which users experience harm, and on which platforms

While most surveys and qualitative research provide detail about which users have been exposed to hazards and experienced harms, this is not provided in platform transparency reports. Knowing who is experiencing harms enables better targeting of interventions. However, there is a challenge in collecting granular information about people, as this will rely to some extent on the user being willing to share information about themselves. Some of which may be sensitive data such as age, ethnicity or vulnerabilities.

*“Our focus is the content that violates our community guidelines, we don’t report on how many people see the content”*

**Social media company**



Similarly, not all measures report on which platform users encounter hazards, which again has implications for future regulation and targeting of interventions. A particular area of concern is ‘hidden’ online spaces—those which are end-to-end encrypted such as WhatsApp—which platforms are unable to monitor or report on. A comprehensive measure of online harm would need to account for hazards and harm occurring in these spaces.

*“Platforms can only report on what they see, their transparency reports don’t account for the encrypted spaces”*

Safety Tech firm

### Most measures ignore positive outcomes

Deciding to only collect data on the negative impacts as a result of a particular online experience could mask potential positive outcomes for people from the same experience. Understanding the scale of positive outcomes—of which there may be none—as a result of exposure to content or experiences online is required to make an informed and conscious decision on whether the mitigation of harm supersedes other outcomes.

### Measures can quickly become out of date—there is a tension between measures which are consistent and stable over time, and updated to reflect emerging trends

Measures that ask people if they have seen particular types of content, or tools that track known hazards, can fast become dated as hazards evolve or new hazards emerge. For example, the types of ‘proana’ (pro-anorexia) content people are engaging with now are likely to be very different to what they will look like in five years’ time. In order to keep up to date with new or unknown harms, more exploratory, ethnographic or qualitative research will be required.

There is a tension between measures that are consistent and stable over time, which are needed to track changes, and measures that can be updated and reflect emerging hazards. It is likely that both will be required.

### Summary:

- This research explored what approaches are currently used to measure the prevalence and impact of different ‘online harms’.
- A wide range of tools and sources currently exist, but do not provide the data required to build a comprehensive or consistent understanding of the prevalence or impact of online harms (overall, or in specific harms areas).
- A wide range of challenges exist, but can broadly be attributed to:
  - Lack of definitions and common understanding of issues
  - Technical limitations of research methods used
  - No consistent measurement and analysis of hazards, risk factors and harms

## What could be?

### *What do we need to know to effectively measure online harms?*

As the section above highlights, current measures are limited in what they are able to claim or show about the quantity and severity of harm caused by different types of online activity, and how it is changing over time. This has immediate implications for how well issues can be addressed and interventions intended to reduce harm evaluated.

Central to understanding online harm is what amount and level of harm can be attributed to online experiences, behaviour or activity. A range of data would need to be collected to accurately determine this—about hazards, risks and harms.

#### **Hypothetical example:**

To be able to say something relatively simple relating to just one area of online harms requires numerous, linked data points. The hazards, risks and harms concept allows for greater clarity in identifying the required data and discussing it. Below is a relatively simple hypothetical example of a research claim that could come from more comprehensive online harms measurement:

**“On Instagram in the past 12 months, teenage users in the UK have seen less content that makes them feel discontented with their bodies, but some teens have been blocked from content they had found supportive.”**

The statement contains information about:

- What platform hazards occur on
- Over what time frame hazards occurred
- What user groups were affected
- Volume of the relevant hazard, and how this has changed
- The direct individual impact of this hazard on users (in this instance, how it made people feel)—both positive and negative

Taking into account the numerous limitations of existing data and measurement approaches outlined in the section above, some key principles can be identified that are prerequisites to being able to effectively measure and track changes in ‘online harms’.

## Online harms measurement must be:

<b>Longitudinal</b> , both short and long term	Not all harms manifest immediately. Measures need to be mindful of the longer-term impacts.
Broken down by <b>user group</b>	<p>While a harm might look negligible at an aggregate level, there may be specific user groups who are suffering, which could be hidden unless sufficient sub-group analysis is possible.</p> <p>Intervention and action may also need to be targeted at specific user groups, an obvious one being children.</p>
Broken down by <b>platform</b>	Harm can occur as a result of hazards across multiple platforms, potentially some more than others. We need to be able to attribute harm to the platform where the hazard originated.
Account for <b>positive</b> and <b>negative</b> impacts	We need to be able to identify the good that can happen and avoid removing this as a result of reducing harm. Without measuring the positives as well as the negatives, action could easily be taken that does more harm than good overall.
Broken down to be <b>UK-specific</b>	To know whether those experiencing online hazards and harm are based in the UK.
Based on both <b>self-report</b> and <b>objective</b> data	<p>Not all hazards and harms can be recognised or articulated by the victim. For example, a child is unlikely to be aware they are being groomed online.</p> <p>Equally, not all hazards and harms are detectable without self-report. For example, some things can cause positive or negative outcomes and you can't tell without asking.</p> <p>So, both self-report and objective data are needed.</p>
Account for <b>unknown</b> or <b>new</b> hazards	We know that if we had done this five years ago, there would be a whole raft of new hazards today that we could not have predicted. Equally, there will be plenty of hazards causing harm right now that we don't yet know about.
Account for <b>'hidden'</b> online spaces	Some hazards can cause harm in areas of the online world not visible to the platforms or outside observers (e.g. encrypted private messages). There is a risk that placing more scrutiny on observable spaces drives more hazards into hidden spaces. Measurement has to account for what happens in these spaces to mitigate this.

## Measuring 'online harm' requires continual development and adaptation

It is extremely unlikely that there will ever be a single measure of all the online harm that occurs.

What can be worked towards is a far more accurate and evidence-based understanding of the prevalence of different hazards and harms, the relationship between them and the role of different risk factors that increase or decrease the chance of harm occurring.

This requires a process for collecting the right data, interpreting it appropriately and using what is learned to inform actions to mitigate harm (and support positive outcomes). As new knowledge and insight about hazards, risk factors and resulting harms (or positive outcomes) emerge, the 'model' should be updated to ensure the use of the most accurate and up to date evidence.

### A model for measuring hazards, risks and harms

A model for measuring online harm and updating understanding with new evidence would have several interlinked components. A cyclical process of data collection and analysis provides the mechanism for identifying and measuring online harm and the key components (hazards and risk factors). The system is also open to new evidence and insight, and the outputs are used to inform interventions. It has these specific components:

1. **Input: Known links between hazards, risk factors and harms.** A set of known or assumed links between hazards, risk factors and harms. Wherever possible this must be informed by insight / evidence, or in cases where this is currently lacking, by logical assumptions of what causes harm. These links inform what data points need to be captured to accurately measure the prevalence and impact of hazards and harms. There is a need for inputs into the data collection and analysis process to come from numerous sources of insight. Otherwise there is a risk they reinforce incorrect measurement—similar to the situation now where the focus is on things we assume to cause harm.

E.g. right now it is assumed (sometimes with good reason and evidence) that hateful language or violent imagery has the potential to cause harm. Identifying this kind of content is an important step in measurement.

2. **Measure causes: Understanding who is exposed to what hazards, and in what contexts.** Informed by the inputs (the known relationships between hazard, risk and harm) it is vital to then measure the prevalence of hazards and related risk factors that are understood to be relevant to specific types of online experience (negative or positive).

E.g. how many people have been exposed to hateful language? Which people, in what context, and how often?

3. **Measure impacts: Degree / severity of negative and positive outcomes from exposure to hazards.** To attribute outcomes to hazards requires an understanding of the impacts on those who have been exposed to them.

E.g. what happened to those exposed to hateful language?

4. **Analyse links: Identifying and evidencing links between hazards, risk factors and outcomes.** Measurement of a hazard and a harm is not evidence that they are inherently linked, or that one causes the other. Analysis is a crucial step as it is where hypotheses are tested and links between experience and outcome are identified. There is not a fixed 'level' for the evidence required to take action, as this depends on the perceived severity of the issue. However, this is a common feature of policy development, which the UK government acknowledges. Analysis can never entirely prove causation, so the precautionary principle may need to be employed<sup>7</sup>. This insight will feed back into the measurement process.

---

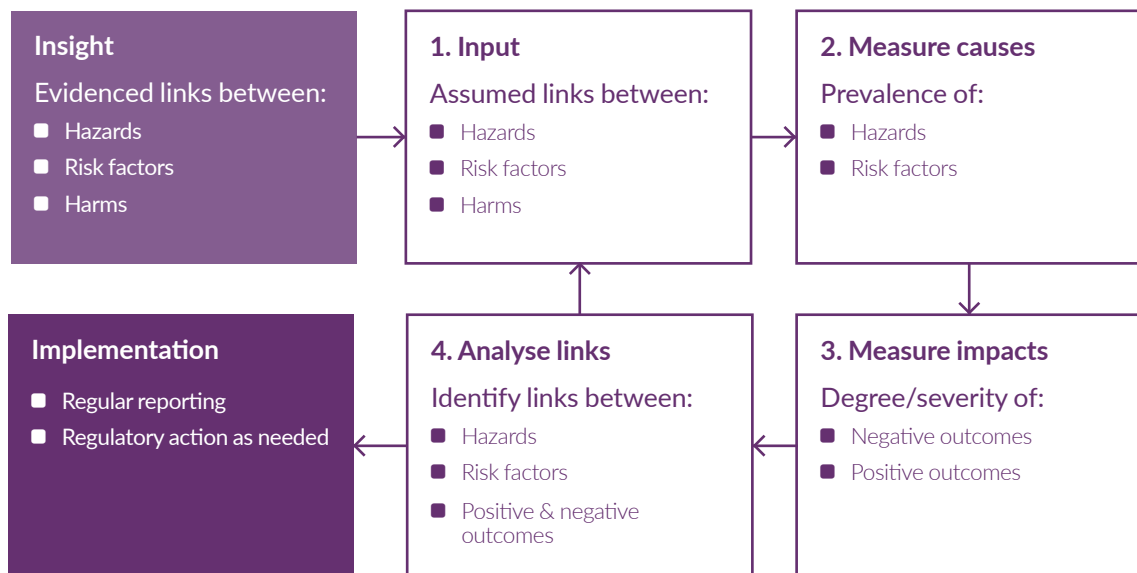
<sup>7</sup> See this [Regulatory Policy Committee guidance note on 'using the precautionary principle' for a useful overview of applying the precautionary principle to policy development.](#)

E.g. is there a link between exposure to a hazard and experience of a harm? Are certain types of people exposed to hateful language more regularly than others, and / or more likely to experience harm? What type of harm is most closely associated with different types of hazard?

**Implementation: Prioritisation of issues and relevant actions / mitigations.** The output of this measurement and analysis cycle can be used to report on the prevalence and impact of ‘online harms’ and, as needed, lead to recommendations or actions intended to mitigate harm. This is a fundamental use of the ‘online harms’ measurement, underpinning Government and regulator action. The effectiveness of these actions would subsequently be evaluated by this broader process.

**Additional insight: New understanding and evidence of links between hazards, risk factors and harms.** This process cannot be a closed system. It needs to be open to new findings and evidence that can be used to improve the accuracy and utility of measurement and analysis. Insight that suggests new links or challenges existing assumptions needs to be an inherent part of this process and is crucial to ensuring a system built on continual learning and improvement.

We’ve drafted a model illustrating what this might look like.



For more detail about how this model can be applied in each harm area, and the potential next steps required, see the harm area-specific reviews in Section 3.

## What does good look like?

This report does not set out in detail every activity that would be required to create the entire process outlined above for each area of online harm. However, there are some general considerations that should be taken into account when developing any part of this process or the research activity that contributes to it.

The table below contains some general principles for any research activities. The list is not exhaustive.

Phase of model	What does good look like?
<p><b>Input:</b></p> <p>Known links between hazards, risk factors and harms.</p>	<ul style="list-style-type: none"> <li>■ Working back from harms, to establish whether they can be attributed (in part or fully) to exposure to online hazards.</li> <li>■ It is widely accepted that establishing true causal links between hazard and harm may not be possible, meaning decisions will likely have to be made about what is considered an appropriate level of evidence to lead to action.</li> <li>■ Drawing on robust evidence to establish links and attribution: Individual cases can identify issues to explore but are not sufficient to prove attribution between hazards and harms.</li> <li>■ The ability to link specific hazards with harms: This would provide more precise information about what hazards look like and what content / behaviours are more closely linked with harm.</li> <li>■ Including information on who and in what context hazards are more likely to cause harm.</li> <li>■ Acknowledging positive / neutral outcomes stemming from the same hazard(s): Amount and severity of harm relative to amount of non-harmful outcomes.</li> </ul>
<p><b>Measure causes:</b></p> <p>Understanding who is exposed to what hazards, and in what contexts. (Hazards)</p>	<ul style="list-style-type: none"> <li>■ Provide an overall picture of how many people in the UK are exposed to different hazards.</li> <li>■ Be as precise as possible: Collecting information in as much detail as possible about all the different types of hazards. This will provide a closer proxy for the prevalence of harm, as different hazards are linked to more or less severe harms.</li> <li>■ Objective and consistent: In order to track the prevalence of hazards over time and across platforms, measures need to be as objective and replicable as possible.</li> <li>■ Determine where people were exposed to hazards (i.e. on what platform—including those which are encrypted).</li> </ul>
<p><b>Measure causes:</b></p> <p>Understanding who is exposed to what hazards, and in what contexts. (Risk factors)</p>	<ul style="list-style-type: none"> <li>■ Account for personal characteristics of users exposed to hazards. E.g. age, gender, ethnicity, previous experiences.</li> <li>■ Account for contextual factors about a user's exposure to hazards, such as the level of exposure a user has had to different hazards, and the combination of hazards they've been exposed to.</li> <li>■ Account for factors that decrease as well as increase risk / likelihood of experiencing harm.</li> <li>■ Is objective and consistent: Could be collected about all people exposed to hazards.</li> <li>■ Account for the level of agency a user has / experiences when interacting with different parts of their online environment.</li> </ul>

Phase of model	What does good look like?
<p><b>Measure impacts:</b></p> <p>Degree / severity of negative and positive outcomes from exposure to hazards.</p>	<ul style="list-style-type: none"> <li>■ Include harms known to be linked to online hazards, as well as the option for users to describe harms not yet known/understood.</li> <li>■ Enable an understanding of how severe harm is.</li> <li>■ Enable an understanding of how likely it is that harm is attributable to exposure to an online hazard.</li> <li>■ Include both harms experienced in the short term and long-term.</li> <li>■ Is linked to their exposure to hazards, and risk factors, described above</li> <li>■ Being on the lookout for and exploring things that are unexplained by our understanding of hazards and risk factors.</li> </ul>

## How far away are we?

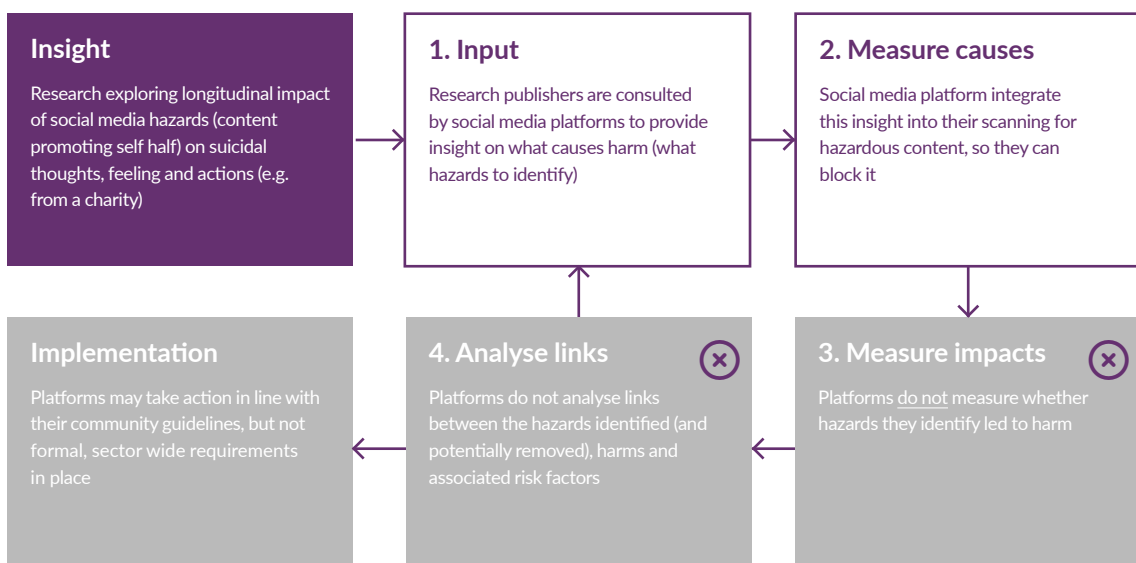
As highlighted already, there is no existing process in place for assessing the total amount of harm that can be attributed to online experiences. Although there are examples of data collection and measurement for most online harm areas, the quantity and quality of data is not consistent.

Some harm areas have data collection activities in place that are closer to collecting data at each stage of this model than others.

### Example 1: Promotion of self-harm

For example, for content which promotes self-harm, there is research exploring the longitudinal impact of different social media hazards (types of content promoting self-harm) on suicidal thoughts, feelings and actions.

This is then shared with platforms who are able to use it to understand more about what types of content can cause harm. Platforms integrate this insight into their scanning for hazardous content, so they can block it.



But that is where it stops. Platforms do not then measure whether these hazards actually caused any concrete harm. In fact, they are likely to just delete them. And there is a concern that some users are then prevented from accessing what they might have found helpful or supportive.

## What would help to bridge the gap between what exists and the 'ideal' measurement model?

To return to the example of Health and Safety in the physical world, we use data collection and measurement at an enormous scale. For instance:

- All workplaces, schools, public spaces collect data
- Health outcomes are monitored, tracked, investigated
- Inspections are carried out
- Materials, buildings, substances are tested
- Data sharing is facilitated and mandated

There are whole sectors of academia, NGOs and private industry dedicated to doing these things.

This enables us to:

- Build an evidence base for how hazards cause harms
- Make the world healthy, not just safe
- Manage hazardous objects, but also the impact of design
- Aim to balance freedom and protection
- Offer more protection to those most at risk – e.g. children, the vulnerable

Measuring 'online harm' will require a collaborative effort between platforms, academia, regulators, government, charities and ultimately, internet users themselves. The government and Ofcom will need to:

- Work with other organisations to shape what data they collect, how it's broken down, and what definitions are used
- Facilitate the sharing of data to the appropriate bodies, whether directly or through transparency reporting
- Produce investigative research into emerging issues, evidencing links between hazards, risk factors and harms
- Track and monitor design changes in the digital world and test their impact on users

### Summary

- Online harm measurement must be: Longitudinal, broken down by user group, broken down by platform, account for positive and negative impacts, broken down to be UK specific, based on both self-report and objective data, account for unknown or new hazards, account for 'hidden' online spaces.
- This report sets out a model for measuring online harms: A cyclical process of data collection and analysis which provides the mechanism for identifying and measuring online harm and the key components (hazards and risk factors).
- Some existing data collection and measurement methods do utilise some elements of the model. However, there is no existing process in place for assessing the total amount of harm that can be attributed to online experiences.
- Successfully measuring online harm will require a collaborative effort between platforms, academia, regulators, government, charities and ultimately, internet users themselves—and a combination of methods and measures.



## SECTION 2:

# Methods currently used to measure online harm

This section contains an overview of the different **methods** that are currently used to measure the prevalence and impact of the harms explored as part of this research, as well as their strengths and limitations. It covers:

- Transparency reports
- Surveys
- Qualitative research
- Automated tools
- Public / government data sets
- Public reporting / helplines

### The difference between a 'method' and a 'measure'

This research explored the **methods or approach** taken to collect data (e.g. a survey with the public, behavioural data from platforms), as well as **specific measures** (e.g. a particular question or set of questions in a survey).

## Transparency reports

A transparency report is a document published by a company, which aims to provide information about processes and practices at that company. In the context of online harms, many of the larger social media platforms regularly and voluntarily publish transparency reports which are publicly available<sup>8</sup>. What platforms choose to share in a transparency report normally relates to their community guidelines.

Community guidelines are the rules developed by platforms which outline what behaviour and content is, or is not, acceptable on their platform. Users of the platform are expected to conform to these guidelines, otherwise they risk having their content removed or their accounts banned.

Some of the largest social media platforms describe their transparency reports in the following ways:

- Facebook describes its transparency report as *“detailing how we are doing at preventing and taking action on content that violates our policies”*<sup>9</sup>
- Twitter explains that the role of their transparency report, alongside the other information they make available, is to *“shine a light on our own practices, including enforcement of the Twitter Rules”*<sup>10</sup>
- Snapchat states *“Our Transparency Report offers important insight into the violating content we enforce against”*<sup>11</sup>
- TikTok uses its transparency report to show *“how we establish and enforce our Community Guidelines”*<sup>12</sup>
- Reddit states: *“We publish this annual report to provide transparency about content that was removed from Reddit, accounts that were suspended, and legal requests we received from third parties to remove content or disclose private user data”*<sup>13</sup>.

The transparency reports often contain information related to the specific online harm areas mentioned in the Online Harms White Paper<sup>14</sup>, as the relevant behaviours or types of content are often in contravention of the stated acceptable use of their platforms, as detailed in their community guidelines or content policies.

---

8 HM Government, 2020. [The Government Report on Transparency Reporting in relation to Online harms](#). [online]

9 Facebook, 2021. [Community Standards Enforcement | Transparency Center](#). [online]

10 Twitter, 2021. [About - Twitter Transparency Center](#). [online]

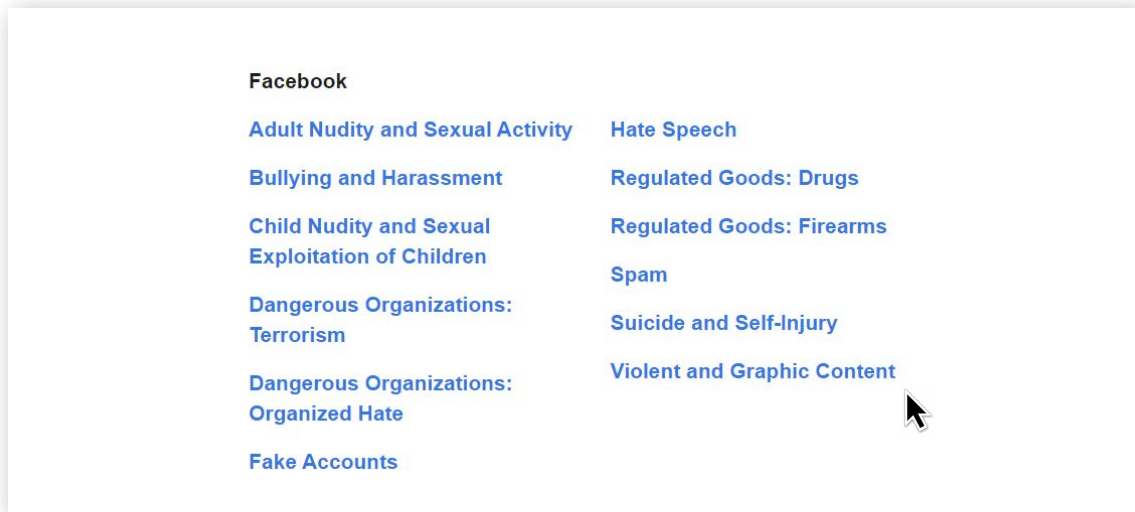
11 Snap Inc., 2021. [Snapchat transparency report](#). [online]

12 Tiktok. 2021. [Tiktok transparency report](#). [online]

13 Reddit, 2021. [Transparency Report 2020 - Reddit](#). [online]

14 GOV.UK. 2020. [Online Harms White Paper](#). [online]

For example, the policy areas that Facebook and Instagram report on (at the time of writing) include<sup>15</sup>:



Transparency reports are created to report on violations of self-defined platform community guidelines and are not created as a measure of online harms. Therefore, it is important to note that while these reports do include data about some of the specific online harms mentioned in the white paper, they have not been created as measurement tools specifically to monitor online hazards and harms; their role is to communicate how well a platform is enforcing its community guidelines.

#### What data can we get through transparency reports:

Transparency reports can tell us about the prevalence of hazards that are found by a platform. The data they collect comes through a number of different channels including user reporting, automated tools that can search terms or some images, and human moderation. Therefore, transparency reports can only tell us what a platform has actually found, which is determined by the policies platforms have in their community guidelines and the tools they use to analyse content.

Currently, transparency reports tend to provide information about the amount of content or accounts removed, rather than user exposure to content. This makes it difficult to understand whether increases in removals are due to better detection of harmful content, or an increase in the prevalence of this content. However, Facebook's Prevalence metric, Google's Violative View Rate, and Snapchat's Violative View Rate are examples of measures that do account for what proportion of content seen by users violated community guidelines.

Examples of the types of online harms data platforms report on in their transparency reports:

- Snapchat records 'turnaround times' which "reflects the median time in hours to action a user report"<sup>11</sup>
- Twitter reported that they suspended 157,815 accounts due to 'Hateful conduct' and removed 1,628,281 pieces of content from July to December 2020<sup>16</sup>
- YouTube reported that 84,777 channels were removed due to 'Harassment and cyberbullying' from January to March 2021<sup>17</sup>, and that their Violative View Rate during this time period was between 0.16 and 0.18% – this metric estimates the "proportion of video views that violate our community guidelines".

Evaluating what transparency reports can tell us according to the hazards, risks and harms framework, it is clear they are able to provide some information about hazards but do not currently provide information in relation to the risk of, or harm caused, by a hazard.

15 Facebook. 2021. [Community standards](#) [online]

16 Twitter. 2021. [Rules Enforcement - Twitter Transparency Center](#). [online]

17 Google/YouTube. 2021. [Google Transparency Report: Youtube Community Guidelines enforcement](#). [online]

### Advantages of transparency reports:

There is already a large degree of alignment between the harms outlined in the Online Harms White Paper and the behaviours and types of content that violate a platform's community guidelines. Therefore, platforms already have an interest in monitoring certain harms on their platforms. Additionally, the reports are easy to access and provide information on how much content platforms detect that violates their guidelines.

Platforms produce the reports regularly and can monitor how the data has changed over time.

While currently the information in the reports is not detailed, the platforms do have access to a large amount of data which informs what can be reported in them. If utilised effectively, this data could be an extremely valuable resource to learn more about online hazards and potentially what makes someone more at risk of coming to harm as a result of a hazard, if data on harms was also captured and analysed. As outlined in the government's report on transparency reporting<sup>8</sup>, developing transparency reports will form a critical element of Online harms regulation.

### Limitations of transparency reports as a measurement tool:

There are currently some overarching limitations stemming what data is used to create the transparency reports and how this is presented, however transparency reports vary from platform to platform and are often updated and adapted over time.

The below broad limitations reflect key issues at the time of writing this report (Mid-2021), acknowledging that many of these limitations could be overcome in future transparency reports.

- Transparency reports are not currently created for measurement, they are created to enforce community guidelines
- They currently predominantly focus on hazards, and can only report about the hazards they find
- There is currently no current standard definition of each harm used across all platforms, so their definitions are based on what violates their community guidelines, making cross platform comparisons challenging
- It is currently hard to track changes in the data over time, as the actions platforms take are based on their community guidelines, which are not fixed. For example, if a platform's definition of cyberbullying was expanded to include new types of content, it is likely that the amount of content removed due to 'cyberbullying' would increase
- The methods for dealing with hazards tend to be removing content or blocking accounts. This does not give any information about whether the hazard stopped, changed or if it migrated to another platform or less visible, encrypted spaces
- Transparency reports currently do not report on how severe a hazard was, how many individuals saw it, or who came into contact with it and how
- There is currently no independent verification of the information that platforms provide and not all platforms provide the same amount of data
- Platforms do not currently report on what happens in encrypted spaces on their sites
- Few transparency reports currently give UK specific breakdowns of data, beyond requests for legal take down of content. However, Snapchat's recent transparency report has done this—a positive step in providing more useful data to a UK regulator.

### What this means for measuring hazards, risks and harms

Overall, transparency reports are a useful indicator of the types and numbers of identified hazards on a platform, but currently do not identify all possible hazards, provide little detail on risk factors and are unlikely to be able to provide useful information on harms. However, given the ability platforms have to collect data, transparency reports have the potential to become an increasingly useful source of information around hazards and risk factors.

### Examples

The table below contains examples of the data collected in current transparency reports across the harm areas explored in this research.

To note, the way platforms have chosen to categorise harms in their measures don't perfectly align with the harm areas we are exploring. For example, platforms reporting on 'Violent and Graphic' content will include content that specifically **promotes** risky or dangerous behaviour, as well as other types of content.

Harm area	Source	Measure / question relating to harm area
<b>Online abuse</b>	Reddit Transparency report 2020 <sup>13</sup>	Reddit removed 51,626 pieces of content due to 'Harassment' and 55,942 due to 'Hateful content' in 2020
	Twitter Transparency report 2020 <sup>10</sup>	Twitter suspended 86,202 accounts due to 'Abuse/Harassment' and removed 1,448,418 pieces of content in the period between July and December 2020
<b>Cyberbullying</b>	Google Transparency Report Q1 2021 <sup>17</sup>	YouTube removed 84,777 channels due to 'Harassment and cyberbullying' from January to March 2021
<b>Access to inappropriate content (pornography)</b>	Google Transparency Report Q1 2021 <sup>17</sup>	1,581,550 videos were removed between January and March 2021 due to 'Nudity or Sexual' content
<b>Content which promotes eating disorders, suicide and self-harm content</b>	Facebook Transparency report Q1 2021 <sup>9</sup>	Five million pieces of 'Suicide and self-injury' content were actioned on in Q1
<b>Content which promotes risky or dangerous behaviour</b>	Facebook Transparency report Q1 2021 <sup>9</sup>	Between 0.03% and 0.04% of views showed content which violated guidelines due to violent and graphic content

## Surveys

Surveys are a research tool used in many different ways. As a research method they provide some overarching challenges and opportunities, which will be relevant to any use case, whether a one-off academic study or an annual online harms-focussed tracking survey. The latter form of surveying does provide some examples of measures that could provide helpful data for assessing the scale of online harms, but also has some specific limitations that are worth noting.

Within a survey, the 'measure' is predominantly a specific question (e.g. X number of people reported Y), or a figure derived from combining responses to several questions (e.g. X number of people reported an experience that met criteria A, B and C).

Surveys present an opportunity to collect self-reported data on people's experiences, behaviours and attitudes. They can be used with anyone capable of comprehending questions and answers, and tailored to specific audiences (e.g. questions written to be age-appropriate). They allow researchers to ask open-ended questions which require respondents to provide answers in their own words, or closed questions asking respondents to choose answers from a pre-populated list. Due to their ability to reach large numbers of people, often relatively easily, surveys are a key tool for quantifying harms, hazards and risk factors.

Surveys are used regularly in one-off and longitudinal research studies around specific areas of online harm, but the most relevant surveys are those that provide regular updates and allow for tracking online harms trends over time. There are several examples of such surveys: Ofcom's Internet Users' Concerns About and Experience of Potential Online harms survey<sup>18</sup>; Ofcom's Adults Media Use and Attitudes survey<sup>19</sup>; Oxford Internet Surveys<sup>20</sup>; ONS Crime Survey for England and Wales<sup>21</sup>.

### Advantages to using surveys as a measurement tool:

Harm is often based on how individuals perceive or experience a hazard. Asking people directly how they felt or whether they experienced harm is often the only way to measure harm using a survey. Whether someone has been bullied, for instance, can be highly subjective and depends to a large extent on how the victim feels as a result of other people's actions.

Depending on sampling, surveys provide a way to identify relative differences between groups of people, based on whatever characteristics are known or collected about respondents. This means surveys can still reveal important information about different groups of people even if they don't comprehensively measure the true extent of a harm. For example, while a survey may not provide the real number of hazards people are exposed to online, it can tell you whether certain people report being exposed to hazards more often than others. These relative differences are incredibly important.

There are also pragmatic reasons why surveys are a valuable tool for exploring online harms:

- They can be relatively easy and inexpensive to run and replicate over time.
- Surveys are highly flexible and can be used to ask about a wide range of hazards and harms, and work towards linking the two. The real limitations are the quality of questions and how many it is feasible to ask a respondent.
- As long as sample is available, they can be used to collect data from specific populations (e.g. a representative survey of adult social media users).

---

18 Ofcom, 2020. [Internet users' concerns about and experience of potential online harms](#). [online]

19 Ofcom, 2021. [Adults' Media Use and Attitudes report. Making Sense of Media](#). Ofcom. [online]

20 OxiS. n.d. [Oxford Internet Surveys - OxiS](#). [online]

21 Office for National Statistics. 2020. [Online bullying in England and Wales - Office for National Statistics](#). [online]

### Limitations to using surveys as a measurement tool:

Surveys rely on self-reported data, which has several critical limitations—some of which can be partly overcome through careful analysis and interpretation of data and critical reflection on what claims can be made—and others which can render it largely ineffective. Key to this is people's (in)ability to accurately recall events, experiences, behaviours, or even how they felt in the past. Other challenges, including social desirability bias, subjective interpretation of questions and concepts, and even dishonest answers, can all put the data, and what can be claimed with it, in doubt. Other limitations include:

- Self-report cannot be used to gauge harms that people don't recognise have happened to them
- Sampling also presents a significant challenge. 'Hard to reach' groups are often underrepresented and challenging to engage via standard sampling approaches, such as using commercial research panels. Reaching truly representative samples is generally extremely difficult and acknowledging the limits and in-built biases of a sample is fundamental to interpreting data accurately
- As noted above, there are only so many questions that a survey can include, so capturing an appropriate level of depth and context around any one online experience will always be challenging

### What this means for measuring hazards, risks and harms

Surveys do provide a route to collecting data on all three aspects, although the scope is limited due to the issues outlined above.

- **Hazards:** surveys can be used to pick up specific issues or content people have come across (whether they considered them to be related to online harms or not)
- **Risk factors:** there are two main routes to understanding risk factors:
  - exploring specific circumstances or factors understood to increase or decrease risk in a survey (i.e. asking specifically about risk factors)
  - through analysis—identifying common characteristics or experiences among groups of people who appear to have better or worse outcomes
- **Harms:** many harms/impacts can be explored via direct questions, and people do not necessarily have to recognise them as harms to respond (e.g. people can report attitudes or behaviours that may indicate longer term harm, without seeing them as such)

### Examples

In the table below we have shown a selection of various survey-based measures that could be used to estimate the prevalence of hazards and harms within certain areas of online harms.

Harm area	Source	Measure / question
<b>Online abuse</b>	Ofcom, Pilot Online Harms Survey (2021) <sup>22</sup>	<p><b>Question:</b></p> <p>Which, if any, of the following have you seen or experienced online in the last four weeks?</p> <p>Answer options include 'Bullying, abusive behaviour or threats'</p> <p><b>Type of claims that can be made:</b></p> <p>6% of internet users report being exposed to bullying, abusive behaviour or threats in the past week</p>
<b>Cyberbullying</b>	ONS, Crime Survey for England and Wales (2020) <sup>23</sup>	<p><b>Question:</b></p> <p>Sometimes, people say or do nasty things to someone. This can happen in person, by phone (texts, calls, video clips), or online (e-mail, instant messaging, social networking, chatrooms). In the last 12 months, have any of these things happened to you?</p> <ol style="list-style-type: none"> <li>1. Nasty messages about you were sent to you</li> <li>2. Nasty messages about you were passed around or posted where others could see</li> <li>3. You were left out or excluded from a group or activity on purpose</li> <li>4. Rumours were spread about you</li> <li>5. Someone called you names, swore at you or insulted you</li> <li>6. Other nasty things happened to you</li> <li>7. None of these</li> <li>8. Don't know</li> <li>9. Don't want to answer</li> </ol> <p><b>Follow-up question:</b></p> <p>Did this happen...</p> <ol style="list-style-type: none"> <li>1. In person</li> <li>2. By a telephone or mobile phone call</li> <li>3. By text message/instant message</li> <li>4. Online</li> <li>5. Some other way</li> <li>6. Don't know</li> <li>7. Don't want to answer</li> </ol> <p><b>Type of claims that can be made:</b></p> <p>"In the year ending March 2020, an estimated one out of five children aged 10 to 15 years in England and Wales experienced at least one type of online bullying behaviour (19%). This equates to approximately 764,000 children."</p>

22 Ofcom. 2021. [Ofcom Pilot Online Harms Survey 2020/21](#)

23 Office for National Statistics. 2020. [Online bullying in England and Wales - Office for National Statistics](#).



Harm area	Source	Measure / question
<b>Cyberbullying</b>	LSE, EU Kids Online (2020) <sup>24</sup>	<p><b>Question:</b></p> <p>Sometimes children or teenagers say or do hurtful or nasty things to someone, and this can often be quite a few times on different days over a period of time. For example, this can include: teasing someone in a way this person does not like; hitting, kicking or pushing someone around; leaving someone out of things.</p> <p>When people are hurtful or nasty to someone in this way, it can happen: face-to-face (in person); by mobile phone (texts, calls, video clips); on the internet (email, instant messaging, social networking, chatrooms).</p> <ul style="list-style-type: none"> <li>■ In the PAST YEAR, has anyone EVER treated you in such a hurtful or nasty way?</li> <li>■ Thinking of the LAST TIME someone treated you in a hurtful or nasty way ONLINE, how did you feel?</li> <li>■ In the PAST YEAR, have you EVER TREATED someone else in a hurtful or nasty way?</li> </ul> <p><b>Type of claims that can be made:</b></p> <p>“In most countries, more than 20% children experienced victimisation”</p> <p>“In the majority of the countries there is no substantial gender difference in victimisation or aggression”</p>
<b>Access to inappropriate content (pornography)</b>	BBFC, Young people, pornography & age verification (2020) <sup>25</sup>	<p><b>Question:</b></p> <p>In the last couple of weeks, have you seen any pictures or videos that would count as pornography?</p> <p>(Yes, No, Prefer not to say, Can't remember)</p> <p><b>Type of claims that can be made:</b></p> <p>“1 in 5 (18%) 11–13 year olds has seen pornography in the past two weeks, rising to 1 in 3 (32%) 14-15 year olds and 4 in 10 (41%) 16-17 year olds.”</p>

24 London School of Economics and Political Science. 2021. [EU Kids Online 2020](#).

25 BBFC, 2020. [Young people, Pornography & Age-verification](#).

Harm area	Source	Measure / question
<b>Promotion of risky/illegal behaviour - self-harm, suicide or eating disorders</b>	Oksanen et al., Young people who access harm-advocating online content: A four-country survey (2016) <sup>26</sup>	<p><b>Question:</b></p> <p>Have you seen the following in the past 12 months? (yes/no answer option):</p> <ol style="list-style-type: none"> <li>1. Sites about ways of physically harming or hurting yourself</li> <li>2. Sites about ways of committing suicide</li> <li>3. Sites about ways to be very thin (e.g. sites relating to eating disorders)</li> </ol> <p><i>Additional questions within this one-off survey study explored happiness and experiences of offline and online victimisation, as well as socio-demographics to determine whether exposure to hazards correlated with other potential risk factors.</i></p> <p><b>Type of claims that can be made:</b></p> <p>“Encountering eating disorder content was more common (17.17%) than encountering self-injury content (10.88%) or suicide content (8.47%)”</p>

<sup>26</sup> Oksanen, A., Näsi, M., Minkkinen, J., Keipi, T., Kaakinen, M. and Räsänen, P., 2016. Young people who access harm-advocating online content: A four-country survey. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 10(2).

## Qualitative research

There is no single approach that defines qualitative research, and the range of methods in this broad category provide many different opportunities for understanding experiences, situations and perceptions in different degrees of depth/detail.

For example, common approaches are:

- **Focus groups:** Putting a group of people together who should have certain experiences or characteristics, to discuss what they think about a topic
- **Interviews:** Exploring people's experiences or perceptions individually, providing an opportunity to delve into both the detail of the topic, and the wider context of their lives, that might help to clarify why someone might have that experience or perception
- **Digital ethnography:** Observing the online behaviour of a certain user or group of users. This can be done in many ways such as:
  - asking users to share their online activity with researchers – e.g. sharing screen record or screenshots of their social media use
  - researchers joining an online community to observe interactions – e.g. researchers join a group or page on a social media platform to observe
  - researchers setting up 'avatars' to replicate the profiles of users with certain characteristics (e.g. a child) and observe the content and interactions they experience. This is particularly valuable when exploring ethically sensitive issues, such as exposure to content relating to self-harm or pornographic content. This method was pioneered in recent research by 5Rights, exploring online harms relating to children and digital design<sup>27</sup>.

Within each method there is a huge amount of scope for variation. Particularly in terms of one-to-one interviews, these can range from high level conversations focussing on people's attitudes, to extremely in-depth explorations of someone's life (where relevant to the core topic).

Qualitative research presents an opportunity to collect:

- Self-reported data on people's experiences, behaviours and attitudes
- Objective data that supports or challenges people's own narrative— for example, what someone has seen on social media
- Contextual information that sheds light on the many factors influencing what people do, experience and think.

Being able to combine these different types of insight to understand individual and collective experiences provides a wealth of opportunities for understanding what is happening and why. Analysing qualitative data provides opportunities to identify links between hazards, harms and risk factors, by identifying various corresponding factors in people's experiences. It is also a useful tool for exploring unknowns, which is why it has such an important role to play in understanding online harm.

---

27 5Rights Foundation, 2021. [Pathways: How digital design puts children at risk](#). [online]

Due to the exploratory nature of qualitative methods and the complexity of many of the online harm areas, qualitative research is a relatively common approach for this subject matter. Some examples of qualitative research studies related to areas of online harm include the qualitative components of Ofcom's Children's and Adult's Media Lives research programmes<sup>28</sup>, and BBFC's Young People and Pornography research<sup>25</sup>.

As with other approaches, qualitative research can be a one-off or used as part of a longitudinal piece of work—tracking people over time and providing, in this instance, an opportunity to identify and explore the long-term impacts of people's online experiences, in the context of other influences on their behaviour.

#### **Advantages to using qualitative research as a measurement tool:**

When it comes to understanding areas of online harms, qualitative research has several key advantages.

Firstly, the in-depth nature of these approaches provides a level of data/insight that is not possible to gain from other methods. As an exploratory tool, qualitative research is vital, and the insight it uncovers can inform other, quantitative approaches to allow insight to be scaled up.

Additional techniques (e.g. validated self-report, behavioural tracking) can provide more objective insights into behaviour. While some of these techniques might be considered quantitative in nature, effective analysis of an individual's behaviour often requires an understanding of the personal circumstances and characteristics of the individual in question, only realistically achievable through qualitative work.

The highly targeted nature of qualitative research, and the relatively small sample sizes required, make it a highly pragmatic way to explore topics and experiences that are challenging for a variety of reasons, such as:

- Low incidence behaviours / experiences
- Hard-to-reach groups
- Complex or sensitive topics

#### **Limitations of using qualitative research as a measurement tool:**

The obvious limitation is that small sample sizes mean that findings are not necessarily representative of the wider population—although efforts can be made to ensure qualitative samples are designed in a way to minimise this risk. Qualitative research can provide an indicative idea of trends and experiences, but should not be relied upon to track changes, or the scale of issues, within the population.

Qualitative methods can be significantly more time and resource intensive than other methods. As the level of detail, depth and quantity of data collected, and analysis increases, so do costs.

As with other approaches, there are challenges engaging certain groups to participate in qualitative research.

#### **What this means for measuring hazards, risks and harms**

The focus of qualitative work is predominantly on exploring what is happening and why. In this context, qualitative work is, and can be, used to understand all three aspects of hazards, risks and harms—particularly in terms understanding the links between them.

It should not be expected to provide insight into the prevalence of different issues—although insight related to scale should be used to challenge quantitative findings where appropriate (e.g. if qualitative work finds that all/most respondents do in fact take part in certain activities, and quantitative research suggests something dramatically different, efforts must be made to understand why discrepancies exist – taking neither at face value).

---

28 Ofcom, 2021. [Children's Media Lives 2020/21](#). [online]

## Examples

The table below contains examples of the data collected in qualitative research across the harm areas explored in this research.

Harm area	Source	Measure / question
<b>Online abuse</b>	Committee on Standards in Public Life, Intimidation in Public Life (2017) <sup>29</sup>	<p>Contains accounts from MPs about the online abuse they have faced.</p> <p>For example, one MP states “It is hard to explain how it makes you feel. It is anonymous people that you’ve never met, true, but it has a genuinely detrimental effect on your mental health. You are constantly thinking about these people and the hatred and bile they are directing towards you.”</p>
<b>Cyberbullying</b>	NSPCC, What children are telling us about bullying (2015/16) <sup>30</sup>	<p>Insights from qualitative data about the type of bullying children have experienced online (and elsewhere).</p> <p>For example: “Young people described malicious and hurtful messages being posted about them on their profiles, blogs, online pictures or posts. The negative messages ranged from bullying and abusive comments about how the young person looked, to directly telling the young person they should go and kill themselves.”</p>
<b>The promotion of eating disorders, self-harm, and suicide</b>	Samaritans and the University of Bristol, Priorities for suicide prevention (2016) <sup>31</sup>	<p>Insights from interviews with over 60 people hospitalised following suicide attempts, to explore their internet use.</p> <p>For example, “For most of those interviewed in the clinical sample, the main purpose for going online was to research methods of suicide, sometimes in great depth. While this did not always lead to action, it made individuals vulnerable by validating their feelings, legitimising suicide as a course of action, and providing knowledge about methods of suicide. Half of those interviewed in the clinical sample planned and carried out a suicide method, based on their online research; some had purchased materials online.”</p>

29 Committee on Standards in Public Life, 2021. [Intimidation in Public Life: A Review by the Committee on Standards in Public Life](#). [online]

30 NSPCC, 2016. [What children are telling us about bullying: Childline bullying report](#). [online]

31 Biddle, L., Derges, J., Gunnell, D., Stace, S. and Morrissey, J., 2016. [Priorities for suicide prevention: balancing the risks and opportunities of internet use](#). [online]

Harm area	Source	Measure / question
<b>Access to inappropriate content (pornography)</b>	BBFC, Young people, pornography & age verification (2020) <sup>25</sup>	<p>Insight from qualitative interviews with young people about their experiences of accessing pornography and the impacts they feel it had on them.</p> <p>It provides examples of the impacts of watching pornography for young people.</p> <p>“Some felt pornography had led to the copying of “rough” sexual behaviour. For example, Lorna’s (18) first boyfriend, whom she entered into a relationship with at age 14, had been pulling her hair and “yanking” her head back during sex. She said that when discussing it with her, he’d told her, “I thought you might like it. The girls in porn like it.””</p>

## Automated tools (Artificial Intelligence, machine learning)

Automated tools, underpinned by data science and artificial intelligence techniques, will likely play an increasingly important role in collecting data and studying various aspects of online harms and hazards. However, it is critical to ensure they are always understood as tools—they need to be designed, maintained and implemented correctly to support measurement.

In many ways, automated tools are hampered by the same fundamental issue that other types of data collection and research are: they can only identify hazards and harms that have already been identified and built into the tools or process.

As highlighted in a recent report for Ofcom on the potential for utilising automated tools to measure some aspects of online harm<sup>32</sup>:

*“Automated tooling has an important and potentially powerful role to play in helping Ofcom to measure online experiences better. However, utilising such tooling effectively is not straightforward: care needs to be given to how to use tools as part of a holistic programme rather than piece-meal; how to extract meaningful and reliable insights with understanding of potential inaccuracies or bias; and how to mitigate the legal and ethical risks associated with mass data collection.”*

As part of our early scoping work, we identified various examples of automated tools being used across the specific online harms areas this research focused on. Automated tools are used widely by social media platforms and Safety Tech firms to identify illegal content (e.g. PhotoDNA – a technology that aids in the removal of child sexual exploitation imagery<sup>33</sup>), or content that violates community guidelines (such as abusive language). ‘Measurement studies’, usually carried out by academics, use automated tools to analyse data sets from social media platforms and provide estimates of what proportion of this data contains ‘harmful content’<sup>34</sup>.

Beyond these uses, there are numerous ongoing efforts to continue to develop and improve automated tools in the detection of a wide range of hazards and even some harms.

### Development of automated technologies and commercial priorities

The main users and developers of automated tools, outside of academia and research, are online platforms and Safety Tech firms (such as Crisp<sup>35</sup>, Factmata<sup>36</sup> and Faculty AI<sup>37</sup>).

While development in this area may align closely with the government’s measurement priorities, it is not guaranteed. Tools that are intended to measure online hazards and harms in a way that meets the needs of DCMS and Ofcom, may not develop organically in the wider commercial market for automated tools. What platforms prioritise for identification and reporting is to some extent a commercial decision, driven by issues that appear to be of most significance or have the greatest potential to cause harm to users and/or subsequently cause reputational damage for the platforms.

Similarly, Safety Tech firms are providing services primarily to private firms with their own commercial interests and priorities—it is these firms who determine, to some extent, what the tools and approaches used by the Safety Tech industry are focused on identifying.

---

32 Faculty, 2021. [Automated Approaches to Measuring Online Experiences](#). [online]

33 Microsoft. n.d. [PhotoDNA | Microsoft](#). [online]

34 Vidgen, B., Margetts, H. and Harris, A., 2019. [How much online abuse is there? A systematic review of evidence for the UK Policy Briefing – Full Report](#). Public Policy Programme Hate Speech: Measures and Counter Measures. The Alan Turing Institute. [online]

35 [Crisp Thinking, 2021](#). [online]

36 [Factmata, 2021](#). [online]

37 [Faculty, 2021](#). [online]

It is likely these interests and measurement priorities will eventually converge in line with UK Government legislation and the requirements of the regulator and relevant firms. But there may well be a need for the development and utilisation of automated tools specifically designed to meet particular measurement challenges, rather than relying on off-the-shelf products or approaches, certainly early on.

Collaboration between government and the Safety Tech industry is already underway and a positive step towards aligning measurement needs with tool development<sup>38</sup>. This includes work with industry bodies such as such as the Online Safety Tech Industry Association (OSTIA)<sup>39</sup>.

This project involved discussion with several Safety Tech firms and academics working with automated tools and data science practitioners and identified the following opportunities and limitations with AI-based approaches.

#### **Advantages of using automated tools as a measurement tool:**

Automated approaches provide a way to conduct studies and identify hazardous content rapidly and at large-scale—with the potential to analyse huge quantities of data (or aspects of the online experiences of whole sections of the population) across different digital spaces.

Tools can be updated and improved as new insight and evidence becomes available. Within the defined parameters they are working in, they should provide consistently accurate results.

There are likely to continue to be rapid developments in the scope and accuracy of automated tools, with new opportunities for bespoke or even ‘off-the-shelf’ products to support the measurement goals. Although it is very difficult to predict when new technologies will be sufficient to meet certain goals, and there will always be limitations that can’t be overcome by improved technology (e.g. access to appropriate data).

Once developed, multiple parties can (in theory) make use of a functioning tool (though effectiveness will vary depending on which platform’s data the AI model was trained on).

Automated tools can be used to identify things at the source (e.g. on the platform where they occur), effectively in real time. This means they are dealing with objective data—the exact content that people are exposed to, or the behaviours exhibited on a platform.

#### **Limitations of using automated tools as a measurement tool:**

One of the biggest challenges for automated tools is the difficulty of interpreting context. As discussed, the context in which a hazard occurs is in many cases an essential element for being able to determine whether harm has occurred, or whether there is a likelihood that harm will occur. This is particularly challenging in areas which are inherently more subjective, such as online abuse or cyberbullying.

For example, in a study on toxic online interaction between adolescents, researchers from the University of Carolina’s AI Institute noted “*Our observations show that individual tweet do not provide sufficient evidence for toxic behaviour, and meaningful use of context in interactions can enable highlighting or exonerating tweet with purported toxicity.*”<sup>40</sup>

Many tools require hazardous content to have already been identified and marked as being hazardous. This is a challenge as the types of content which are considered to potentially cause harm are constantly evolving. Models are time-sensitive and may struggle to keep up with the latest trends.

AI/machine learning tools need to be trained on suitable data. This adds an additional layer to the process as the generation of training data sets is a key step for the development of certain types of tools.

Currently, automated tools tend to skew towards analysing language, rather than images or video content. This is essentially due to the difficulty of successfully analysing these more complex stimuli.

---

38 GOV.UK. 2021. [The UK Safety Tech Sector: 2021 Analysis](#). [online]

39 [Ostia, 2021](#). [online]

40 Wijesiriwardene, T., Inan, H., Kursuncu, U., Gaur, M. Shalin, V. L., Thirunarayan, K., Sheth, A., & Aprinar, B. 2020.: [ALONE: A dataset for toxic behavior among adolescents on Twitter](#). [online]



### What this means for measuring hazards, risks and harms

Automated tools are predominantly focused on identifying hazards. While risk factors and harm are not beyond scope for identification by automated tools, the challenges are significant due to the contextual and subjective nature of some of these aspects, as well as limited access to appropriate contextual data (e.g. personal information about users).

### Examples

In the table below we have shown a selection of various automated approaches that could be used to estimate the prevalence of hazards within certain areas of online harms.

Harm area	Source	Measure / question
<b>Online abuse</b>	Farrell et al., MP Twitter Engagement and Abuse Post-first COVID-19 Lockdown in the UK: White Paper (2021) <sup>41</sup>	<p>This work analyses Twitter abuse in replies to UK MPs in the period of June to December 2020.</p> <p>It uses an automatic abuse detection method, identifying terms or short phrases that had been previously classified as abusive.</p> <p>The authors note that this method underestimates the amount of abuse, as although their automatic tool was able to find more obvious verbal abuse, it missed linguistically subtler examples.</p> <p>In June 2020 4.4% of all replies to MPs' tweet were 'abusive'.</p>
<b>Cyberbullying</b>	Chatzakou et al., Mean Birds: Detecting Aggression and Bullying on Twitter (2017) <sup>42</sup>	<p>Twitter users were labelled as normal, aggressive, bullying, or spammer by people manually analysing batches of their tweet. Features of the user and their tweet were extracted and analysed to determine the characteristics more common to those classified as 'bullies'.</p> <p>Machine learning classification algorithms can then accurately detect users exhibiting bullying and aggressive behaviour (over 90% accuracy) based on this dataset.</p>

41 Bontcheva, K., Bakir, M. and Farrell, T., 2021. [MP Twitter Engagement and Abuse Post-first COVID-19 Lockdown in the UK: White Paper](#). Department of Computer Science, Sheffield University. [online]

42 Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G. and Vakali, A., 2017. [Mean Birds: Detecting Aggression and Bullying on Twitter](#). Proceedings of the 2017 ACM on Web Science Conference., [online]

Harm area	Source	Measure / question
<b>Promotion of risky/illegal behaviour - self-harm, suicide or eating disorders</b>	Scherr et al., Detecting Intentional Self-Harm on Instagram: Development, Testing, and Validation of an Automatic Image-Recognition Algorithm to Discover Cutting-Related Posts (2019) <sup>43</sup>	An image-recognition algorithm was used to explore the relative prevalence of 'non-suicidal self-injury' such as cutting, in pictures posted within 48 hours on Instagram under #cutting (n = 4,219) and #suicide (n = 7,910)
<b>Access to inappropriate content (pornography)</b>	Perez et al., Video pornography detection through deep learning techniques and motion information (2017) <sup>44</sup>	Automated pornographic detection was explored combining static (picture) and dynamic (motion) information.  The best proposed method was tested on a dataset of 800 challenging test cases, and was able to accurately classify 97.9% of cases.

43 Scherr, S., Arendt, F., Frissen, T. and Oramas M, J., 2019. Detecting Intentional Self-Harm on Instagram: Development, Testing, and Validation of an Automatic Image-Recognition Algorithm to Discover Cutting-Related Posts. *Social Science Computer Review*, 38(6), pp.673-685

44 Perez, M., Avila, S., Moreira, D., Moraes, D., Testoni, V., Valle, E., Goldenstein, S. and Rocha, A., 2017. Video pornography detection through deep learning techniques and motion information. *Neurocomputing*, 230, pp.279-293.

## Public / government data sets

Data sets collected by government and other public bodies may provide data that demonstrates the scale of some online harms or hazards, despite not having been created with this intention. Across government departments data is collected for a range of other functions. Some of this data may provide useful proxies for some of the outcomes associated with online behaviour and experiences, that can be tracked over time.

For example, Police data on hate crimes<sup>45</sup> potentially provides a valuable route to understanding the scale of online hate occurring, via the proxy of how many instances are reported to the Police. The ONS holds data on suicides<sup>46</sup> and the NHS records data on hospital admissions for intentional self-harm<sup>47</sup>, which could be used as one piece of evidence, alongside other data, relating to people's online experiences of content that promotes self-harm and suicide (acknowledging the extreme caution that would need to be used if referring to such datasets to make inferences about the relationships with online experiences).

### Advantages of using public data sets as a measurement tool:

One of the key advantages with public / government data sets is that they are often consistent year on year—in many instances their core function is tracking behaviours of interest within the population over time.

In some cases—e.g. crime figures—the data may contain valuable context such as the characteristics of victim/person reporting.

On a pragmatic level, these data sets already exist, have mechanisms in place for collecting and analysing the data and have historical data that can be re-assessed if required.

### Limitations of using public data sets as a measurement tool:

There are several key limitations of this kind of data. Firstly, many existing data sets do not attribute the experience/behaviour they are collecting to online activities or experiences, such as data on hospital admissions due to self-harm. They are only able, therefore, to play a limited role in assessing the scale of online harm, acting as a tertiary piece of evidence that may indicate a trend that needs to be acknowledged.

What these sources can tell us is informed by the level of evidence available on the links between online behaviour and certain outcomes (e.g. the greater the evidence for a link between content that promotes self-harm and self-harm, the more important overarching trends in self-harm are).

Many existing data sets that could be of use are the result of official reports. As such, they are focussed on more severe, illegal harms, and far less on the 'legal but harmful' harm areas.

Official statistics are often subject to a significant time delay, meaning they are likely not to be suitable for in the moment analysis.

### What this means for measuring hazards, risks and harms

Predominantly, any existing data sets will focus on specific harms that have occurred. Usually, the harms will be particularly severe to warrant existing data collection.

---

45 GOV.UK. 2020. [Hate crime, England and Wales, 2019 to 2020](#). [online]

46 Office for National Statistics. 2019. [Suicides in England and Wales - Office for National Statistics](#). [online]

47 NHS Digital. 2020. [Hospital admissions for intentional self-poisoning and self-harm - NHS Digital](#). [online]

### Examples

In the table below we have shown a selection of various public / government data sets that could be used to estimate the prevalence of hazards within certain areas of online harms.

Harm area	Source	Measure / question
<b>Online abuse</b>	Home Office, Hate Crime data (2017/18) <sup>48</sup>	In 2017/18, two per cent (1,605 offences) of all hate crime offences were indicated as having an online element in the year ending March 2018.
<b>Promotion of risky/ illegal behaviour - self-harm, suicide or eating disorders</b>	Public health England Fingertips data (2019/20) <sup>49</sup>	There were 664.7 people admitted to hospital per 100,000 people as a result of self-harm, for those aged 15-19 years old in England.

48 GOV.UK. 2018. [Hate crime, England and Wales, 2017 to 2018](#). [online]

49 PHE, P., 2021. [Self Harm](#). [online] Public Health

## Public reporting sites and helplines

There are various dedicated websites or helplines where members of the public can report incidents of online harm or hazardous content, often with a specific focus, and/or receive support for specific issues. Both are able to provide data that could allow insights into the prevalence of certain hazards or harms.

Reporting sites provide an opportunity for members of the public to formally report instances of harm they, or someone they know, has experienced online. Examples include: Tell Mama<sup>50</sup>, a platform for people to report Islamophobia; Community Security Trust (CST<sup>51</sup>), which provides a platform to report antisemitic incidents; and the UK Safer Internet Centre, which provides the option to report a wide range of harmful content on the Report Harmful Content platform<sup>52</sup>.

Other platforms/services where people receive support for a particular issue can also provide an insight into the prevalence of that issue. Use of the Samaritans helpline or Mind services, for example, could provide indicative data on the rates of people experiencing mental health issues; while uses of Childline could provide data on children's experiences of abuse. However, for these sources to be of particular use, they would need to be able to attribute the experiences service users are relaying to them to their online experiences, which many are not set up to do.

### Advantages of using public reporting sites and helplines as a measurement tool:

These sites and services already provide relatively regular updates on the number of reports and interactions they have, as well as any other information they collect.

While the information they report on is generally limited, there are opportunities for these platforms to gather data on the characteristics of those making reports and using their services (where appropriate) as well as additional information about incidents in question, especially the role of digital/online.

### Limitations of using public reporting sites and helplines as a measurement tool:

There is a relatively high threshold for something to be reported to a third party, or for someone to seek support. Therefore, the reports and interactions these platforms are able to provide data on may only represent a specific subset of the kinds of experiences or behaviours of interest—they require that someone had recognised that a harm has occurred or has identified a hazard.

Because use of public reporting sites and helplines requires both an awareness of them, and the capability and motivation to use them, the reports made are likely to be a significant underrepresentation of all the related harms or hazards occurring at any one time.

It is also difficult to attribute changes in the number of reports to any specific causes. For instance, increased reporting could come from increased awareness of a reporting site rather than an increase in the incidents of interest—interpreting this change as anything else could be inaccurate or misleading.

While there are opportunities to collect more relevant information about the online nature of people's experiences, there are also limits to how much data helplines and reporting sites can collect. For example, if someone has called a helpline to seek support for self-harm, it is unlikely to be appropriate to try and collect data about online behaviours.

---

50 Tell MAMA., 2018. [Normalising Hatred: Tell MAMA Annual Report 2018](#). [online]

51 [CST, 2021](#). [online]

52 Report Harmful Content, 2021. [Submit a Report of Harmful Content](#). [online]

### What this means for measuring hazards, risks and harms

Reporting sites capture both hazards and harms, largely depending on the role and area of interest of the site. A site that captures reports of 'harmful content' is predominantly capturing instances of people coming across hazards, while a site that reports experiences of—for example—islamophobia is capturing people's perceived experience of that harm.

Helplines currently collect data related to experiences of harms; however, the link isn't always able to be made between harms and online hazards.

### Examples

In the table below we have shown a selection of various public reporting sites and helplines that could be used to estimate the prevalence of hazards within certain areas of online harms.

Harm area	Source	Measure / question
Online abuse	Tell Mama, Normalising hatred (2018) <sup>50</sup>	There were 327 verified reports of anti-Muslim attacks online reported to Tell Mama, down 10% from the figure of 362 verified reports in 2017.
Cyberbullying	NSPCC, Childline bullying report (2018/19) <sup>53</sup>	There were 15,851 counselling sessions with apeer bullying, which encompasses both face-to-face and online).

.....  
 53 NSPCC, 2018. [Childline Annual Review](#). NSPCC. [online]

## SECTION 3:

# Measurement in the four harm areas

This section is concerned with some of the specific areas of online harm of particular interest to DCMS, as noted in the Online Harms White Paper.

The harms areas discussed in detail here are:

- Online abuse
- Cyberbullying
- Access to inappropriate content (pornography)
- Promotion of risky / illegal / dangerous behaviour – notably the promotion of eating disorders, self-harm and suicide

Online abuse was explored in relation to adults, while cyberbullying, access to inappropriate content and the promotion of risky/illegal/dangerous behaviour were explored in relation to children.

For each harm area, we discuss key definitions and provide a short summary of existing measures for assessing their prevalence and impact. Potential 'next steps' for measurement are also provided, focused on the immediate challenge of providing a usable benchmark for understanding the scale and impact of each harm.

**A note on recommendations specific to the harms areas reviewed in this document**

For each of the harms discussed, the limitations of the current measures are outlined, as well as some potential improvements that could be actioned in the short-term to mitigate / overcome them.

Given the fundamental limitations associated with each of the methods currently used to measure online harms, outlined in [Section 2](#), multiple methods will be required to improve the accuracy of online harm measures. Recommendations must be read with these limitations in mind.

The recommendations are linked to specific parts of the wider online harms measurement process outlined in [Section 1](#).



## A. Online abuse

### Defining online abuse

Existing research usually separates online abuse into two main types<sup>34</sup>:

- Abuse directed against a group, usually called 'hate speech'
  - "[Hate speech] broadly includes negative textual, visual or audio-based rhetoric that attacks, abuses, insults, harasses, intimidates, and incites discrimination or violence against an individual or group due to their race, ethnicity, gender, religion, sexual orientation or disability" (Davidson et al., 2019<sup>54</sup>)
- Abuse directed against an individual, usually called 'harassment' or 'cyberbullying'
  - A definition for online (and not necessarily illegal) harassment is: "aggressive, intentional acts carried out by a group or individual, using electronic forms of contact against an individual" (Dadvar et al., 2013 as cited in Vidgen et al., 2019<sup>34</sup>)

When talking about measures of online abuse this is likely to include both 'online harassment' and 'online hate', as it can be difficult to distinguish between the two.

The definitions above provide a useful summary, but there are no universally accepted or consistently used definitions for online hate speech or online harassment—as highlighted in a Rapid Evidence Assessment of Adult Online Hate, Harassment and Abuse<sup>54</sup> and the Turing Institute's review into the prevalence of Online Abuse<sup>34</sup>. Definitions tend to focus on describing different types of abusive behaviour and the intent behind this behaviour (e.g. to intimidate or insult).

The main types of online harassment encountered according to Davison et al., 2019<sup>54</sup> are:

- Offensive name calling
- Purposeful embarrassment
- Physical Threats
- Sustained Harassment
- Stalking
- Sexual Harassment

Specific tactics may be used to carry out online abuse such as pile-on harassment.

**'Pile-on' harassment:** "Pile-on harassment occurs when many individuals, acting separately, send messages that are harassing in nature to a victim." "Pile-on harassment is a form of group harassment in which a number of individuals each send messages that, when taken together, cause alarm or distress – even though, taken individually, no message reaches a criminal threshold. This form of harassment is often, though not always, targeted at high profile individuals and can have a devastating impact." This report notes that pile-on harassment is likely encouraged by the **disinhibition effect** and the perception of **anonymity** (Law Commission<sup>55</sup>).

54 Davidson, J., Livingstone, S., Jenkins, S., Gekoski, A., Choak, C., Ike, T. & Phillips, K. (2019). [Adult Online Hate, Harassment and Abuse: A Rapid Evidence Assessment](#). UK Council for Child Internet Safety.

55 Law Commission, 2020. [Harmful Online Communications: The Criminal Offences - A Consultation Paper](#). [online]

## Challenges in defining online abuse

Online abuse is a broad category containing many different types of behaviour, likely to have a different outcome depending on the severity of the abuse, type of abuse, and the recipient of the abusive content.

Definitions of online abuse are subjective to some extent and often dependent on interpretations of the abuser's intent and behaviour. What can be considered 'threatening' or 'embarrassing' or 'attacking' will vary from person to person and is highly dependent on the personal characteristics of the recipient, as well as the context in which it occurs.

## Summary of the current measures of online abuse

Measures in this area often lack precision about what types of content or behaviour people have been exposed to, that fall under the category of 'online abuse' or 'online harassment'. For example, most surveys ask people if someone has 'experienced abuse, harassment, or hate online'<sup>56, 57</sup>, with little detail about what this may have entailed.

### Prevalence of online abuse

Much of the research (i.e. surveys, measurement studies, qualitative research) conducted into online abuse is done on a one-off basis, and therefore cannot track changes over time.

The Ofcom Internet User's Experience of Potential Online Harms<sup>18</sup> survey (which has now been replaced by the Ofcom Pilot Online Harms survey<sup>22</sup>) is one of the few annual surveys covering online abuse. It also provides one of the most detailed breakdowns of specific hazards within this area, though some hazards such as 'trolling' and 'offensive language' included in the survey are likely to be interpreted differently by different users. Interpreting something as 'trolling', for instance, also depends on various contextual factors and the personal characteristics of the recipient. The survey aims to include a representative sample of UK adults, with quotas for age, gender, location and social grade. However, it does not have quotas, or provide breakdowns in the data, for specific characteristics which may be linked to abuse such as sexuality and disability, and is unlikely to include people who have a more public online profile (e.g. politicians, celebrities).

There are some populations who have been highlighted as being more likely to experience online abuse, for example, those with disabilities and those from ethnic minorities<sup>34</sup>. Some of the measures focus on specific groups' experiences of online abuse. For example, public reporting sites, such as Tell Mama – for people to report Islamophobic content or abuse, and surveys run by charities focusing on supporting certain groups, such as Galop's survey of 700 LGBT+ people's experiences of homophobic abuse.

### Impact of online abuse

While there is some qualitative and quantitative data on the impact of online abuse<sup>54, 58</sup>, this is often not linked to precise information about what the user has been exposed to.

For example, Amnesty International research into the impact of online abuse against women<sup>59</sup> asked people if they've experienced "online abuse or harassment", and how it made them "feel or act". This makes it hard to understand how specific experiences (e.g. repeated exposure to a certain type of behaviour on a certain platform) were related to harmful outcomes.

---

56 Amnesty International and Ipsos MORI., 2017. [Poll: Online abuse or harassment against women – Online experience](#). [online]

57 Thomas, K., Akhawe, D., Bailey, M., Boneh, D., Bursztein, E., Consolvo, S., Dell, N., Durumeric, Z., Kelley, P.G., Kumar, D., McCoy, D., Meiklejohn, S., Ristenpart, T., & Stringhini, G. (2020). *SoK: Hate, Harassment, and the Changing Landscape of Online Abuse*. [online] Available at: sok-abuse.pdf (cornell.edu)

58 Galop, 2020. [Online Hate crime Report 2020 - Challenging online homophobia, biphobia and transphobia](#). [online]

59 Amnesty International. 2021. *Toxic Twitter - The Psychological Harms of Violence and Abuse Against Women Online*. [online]

There are only very limited attempts to measure the severity of the abuse people are exposed to. This makes it very difficult to determine whether there are certain types of abusive content and behaviour that contribute to the harm people experience to a greater degree than other types.

### Use of AI in online abuse detection

AI used by platforms and academic measurement studies to detect online abuse tends to be limited. AI models often focus on identifying specific words or phrases, deemed to be likely to cause offense, but are unable to identify more subtle forms of abuse. For example, one Reddit user set up the community 'r/blackfathers' and ensured that no content was able to be posted in it by continuing to delete posts. This meant that users who searched for the subreddit were directed to a message telling them there was "nothing there", perpetuating a derogatory stereotype. This less overt form of hate is unlikely to be picked up by AI which is predominantly trained to focus on abusive language. The systematic review of online abuse measurement by the Alan Turing Institute identified key challenges including non-representative data sets or samples; lack of UK focus; and high error levels when deploying tools outside test environments.<sup>34</sup>

*"[Social media interactions between adolescents] exhibit complex linguistic and contextual characteristics, making recognition of such narratives challenging...Our observations show that individual tweet do not provide sufficient evidence for toxic behaviour, and meaningful use of context in interactions can enable highlighting or exonerating tweet with purported toxicity"*

Wijesiriwardene, et al., 2020<sup>40</sup>

### Governmental data

Government statistics on online hate crime have been published in the past with "experimental" figures reported in 2017/18, from 30 out of 44 police forces, showing that 1605 online hate crimes were recorded in England and Wales which accounts for around 2% of all hate crimes<sup>60</sup>. While these figures may be used as an indicator of the amount of online abuse occurring, they have not been reported since. There are also numerous factors that could influence reporting rates (e.g. changes to reporting procedures), meaning data is unlikely to be consistent.

## Key limitations and potential next steps

Based on the existing limitations of research and data around online abuse, there are some immediate next steps / opportunities for improving the underlying evidence base as well as the kind of data we collect that indicates how the prevalence and impact of this specific type of online harm is changing over time.

We have outlined these in the table below. Each is linked to one of the key parts of the overarching measurement process outlined in Section 1.

---

60 GOV.UK, 2018. [Hate Crime, England and Wales 2017/18](#).

Which aspect of the measurement process does it relate to?	Key limitations of existing measures	Recommended next steps / opportunities
<p><b>1. Input.</b></p> <p>A set of assumed links between hazards, risk factors and harms.</p>	<p>Limited research exploring how a range of harmful outcomes are associated with different forms of online abuse.</p>	<p>Longitudinal research exploring the links between a range of online abuse and harmful outcomes. This would enable better targeting of interventions, focusing on the types of online abuse which appear to cause the greatest harm.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Commission a new longitudinal study that explores exposure to online abuse.</li> <li>■ Include an exploration of online abuse in existing longitudinal work. E.g. any existing cohort studies.</li> </ul>
<p><b>2. Measure causes (hazards) and risk factors.</b></p> <p>Measure the prevalence of hazards and the risk factors involved.</p>	<p>Surveys often ask about whether users have seen or been exposed to “online abuse”, “harassment” or “hate” with limited detail on what this involved – e.g. how often, how severe, how did it occur?</p>	<p>Surveys should aim to gather more granular detail on the type of abuse, harassment, and hate people are experiencing, to establish where certain types of content are linked to more severe outcomes. E.g. understanding what type of abuse it was, whether it was overtly offensive or highly targeted at an individual, what made it offensive/abusive etc.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Commission new quantitative research with the specific aim of providing a greater level of detail on experiences of online abuse.</li> <li>■ Provide recommendations on a question set and/or expectations for the types and granularity of data surveys covering online abuse will report on to be valuable to the overall online harms measurement process.</li> </ul>
	<p>Similarly, platforms often report how much content has been removed due to “abuse/harassment” as opposed to more granular reasons. This makes it challenging to track whether certain types of abusive content / behaviour have increased while others may have decreased.</p>	<p>Platforms should aim to report more detailed information about the types of ‘abuse / harassment / hate’ that was exposed to users on their platform. E.g. what type of abuse it was, whether it was overtly offensive or highly targeted at an individual, what made it offensive/abusive, what format of content it was (photo, text etc.).</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Set a minimum requirement and expectation for what level of granularity platforms will report on (privately and/or publicly).</li> <li>■ Outline an expected methodology for use by platforms and/or a set of clear research objectives that their chosen methodology must meet.</li> </ul>

Inconsistent definitions of 'online abuse/harassment' make it hard to compare data across studies and platform reports.

Online platforms should work from more consistent definitions of online abuse, hate and harassment.

**Potential actions:**

- Clarify for key stakeholder audiences (researchers, platforms etc.) the different types of online abuse, with clear examples, so that a wide range of audiences can measure the same thing (or have better sight of where different things are being measured and potentially conflated).
- Ensure these definitions are updated and it is generally accepted which aspects are more or less difficult to ascertain (providing some expectations or baselines for what relevant parties should be measuring as a minimum).

It is not possible to identify instances where high volumes of online abuse are targeted at one individual from current platform reporting.

Platform data / measurement studies should examine instances where a high volume of abuse / hate is targeted at an individual.

**Potential actions:**

- Include this within any platform-reporting expectations/briefs.

There are significant challenges for automated tools (e.g. AI) to detect context around potential abusive content.

Continual work is needed within AI to develop models more sensitive to context.

Research and platform data which provides a better understanding of what online abuse looks like should feed into developing new AI models.

**Potential actions:**

- Provide a platform or incentive for different actors working in this field to share learning and data that can help others improve their own tools or research with AI / machine learning.
- Incentivise (or compel) platforms to provide suitable test/ learning data sets to relevant actors.

	<p>There is limited understanding of the impacts of online abuse on different audiences.</p>	<p>More qualitative / observational research is needed to understand the impacts on people and identify how different forms of online abuse affect different people.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Ensure qualitative research collects extremely granular detail on the content people are exposed to (potentially working with platforms to access objective data on content exposure and respondent's online activity).</li> <li>■ Ensure a suitably diverse sample is included in any key longitudinal or qualitative research into the impacts of online abuse (whether expanding existing projects or starting new work).</li> </ul>
	<p>There is limited understanding of the impacts of online abuse based on the context of abuse, such as the volume of content someone is exposed to over time or where on the platform abuse is seen.</p>	<p>Platform-based research or reporting must account for a range of contextual risk factors.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Include this within any platform reporting expectations/ briefs.</li> </ul>
<p><b>3. Measure impacts.</b></p> <p>Measure what actually happens to people / society who have experienced the hazards.</p>	<p>Surveys do not consistently collect data on the impact of experiencing online abuse / harassment / hate.</p>	<p>Surveys should ask about the impact of exposure to abusive behaviour and content online, to better understand the prevalence of harm.</p> <p>Asking people about the outcomes of online abuse needs to be done in an ethical way, ensuring that respondents who may have experienced harm are safeguarded and supported.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Develop a set of questions to better understand the impact of abusive behaviour, both at a personal and societal level, with careful ethical considerations</li> </ul>
	<p>Qualitative research into the impacts of online abuse is limited</p>	<p>More qualitative research is required to understand the impact of online abuse on a range of people.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Carry out qualitative research with groups who have been identified as more likely to be exposed to online abuse, as well as wider audiences</li> </ul>

## B. Access to inappropriate content (specifically pornography)

### Defining access to inappropriate content

'Inappropriate content' has an extremely broad definition, with what counts as 'inappropriate' determined by societal and personal factors.

It has been defined variously by organisations with a focus on children's online safety as content that is not suitable for someone's age, and that may upset or worry them (NSPCC).<sup>61</sup> It can also be said to include content—defined as images or information—that is directed at adults; inaccurate information or "information that might lead or tempt your child into unlawful or dangerous behaviour" (Internet Matters)<sup>62</sup>. Definitions of inappropriate content often classify content such as "pornography, anything encouraging swearing, vandalism, racism, bullying, terrorism or suicide, violent content, animal cruelty and sites encouraging gambling" (BT).<sup>63</sup>

It includes content that may be covered by other types of 'online harm', so there is a lot of crossover between the ideas of inappropriate content, risky or dangerous behaviour and specific types of content such as self-harm, suicide or anorexia content.

Defining something as inappropriate based on age and/or the potential that it may cause someone direct or indirect harm means what we categorise as inappropriate is informed by societal factors, e.g. what we consider at a societal level should be age restricted, such as pornography or violent films and games, and by the characteristics of the person being exposed to it, e.g. their age, gender, whether they are likely to be influenced more or less by things they see online.

#### Access to inappropriate content – specifically pornography

Pornography is a key form of inappropriate content identified in the Online Harms White Paper and the focus of this particular review of measures. The Crown Prosecution Service definition of pornography is content that is "of such a nature that it must reasonably be assumed to have been produced solely or principally for the purpose of sexual arousal".<sup>64</sup>

What counts as pornography is not as subjective a question as what counts as, for instance, online abuse. However, when trying to understand the impact of exposure to pornography on under 18s, it's vital we recognise the immense diversity of pornographic content that exists online—some of which we might reasonably assume is likely to be inherently more harmful or impactful on the attitudes and behaviour of viewers.

---

61 NSPCC. 2021. [Inappropriate or explicit content](#). [online]

62 Internet Matters. 2021. [What parents need to know about inappropriate content](#) | Internet Matters. [online]

63 BT. 2021. [Inappropriate Content and How to Spot It](#). [online]

64 CPS. 2021. [Extreme Pornography - Legal Guidance, Sexual offences](#). [online]

This definitional challenge has significant consequences, as treating all pornography as equal for the purposes of measuring online harm is likely to create significant 'noise' in the data. At the aggregate level, without suitable differentiation between types of content and patterns of viewing behaviour, we would miss instances where the majority of 'harm' is actually caused by only a certain type of content. Treating all pornography as equal flattens our understanding and leaves us with only blunt, inaccurate tools (e.g. removing all pornography, rather than harmful pornography) to mitigate issues

Please note that we are concerned here specifically with pornography that is legal for those 18+. Obscene pornographic material and child sexual abuse imagery are illegal forms of content, and the identification, removal and the ramifications for viewers and providers of this content are covered by a different legal process.

## Summary of the current measures of access to inappropriate content (pornography)

Measures in this area are limited to self-report studies in which children and adolescents are asked specifically about their exposure to pornography.

All major studies identified in the literature review that explore pornography access among children in detail, appear to be one-off pieces of research, providing an immediate challenge/limitation in terms of observing changes over time.

One of the largest dedicated studies on this topic in the UK was a qualitative and quantitative piece of research conducted on behalf of the BBFC in 2019<sup>65</sup>. Even with exploratory qualitative research and a large-scale quantitative survey with children, the accuracy and precision of the data was limited and focused on overarching trends and experiences rather than the direct link between hazards and potential harm.

None of these studies adequately address the issues of the type (possibly 'severity') of content and patterns of use/exposure.

### Using AI to detect pornography

AI has been developed that can detect pornographic images and videos in real time, with the intention of blocking flagged content for underage users.<sup>65</sup> Other systems have been developed that can estimate someone's age, and have been used to identify explicit content containing underage individuals (e.g. in cases where under 18s are selling sexually explicit images or videos of themselves).<sup>66, 67</sup>

While these technologies provide opportunities to better identify instances of underage exposure to pornographic content, neither is being used systematically to collect this data and provide an estimate of the prevalence of the relevant hazard.

---

65 Perez, M., Avila, S., Moreira, D., Moraes, D., Testoni, V., Valle, E., Goldenstein, S. and Rocha, A., 2017. [Video pornography detection through deep learning techniques and motion information](#). *Neurocomputing*, 230, pp.279-293.

66 Yoti, 2021. [Yoti Age Scan - White Paper Full version](#). [online]

67 Biometric Update. 2020. [Yoti AI age estimation used in BBC investigation of underage porn on social media](#) | Biometric Update. [online]



**Linking access to pornography to harms**

There is limited evidence explicitly linking underage access to pornography to short- or long-term harms. However, studies have identified various related findings, such as perceptions that pornography has altered people’s expectations of sex and healthy sexual relationships (at all ages), people’s behaviours and sexual preferences and that at least some people have seen pornographic content that has upset or disturbed them. There is also qualitative evidence that in more extreme cases, early exposure to pornography—particularly forms that are violent or demeaning—may have long lasting effects on someone’s personal experiences with sex as they get older.

**Key limitations and potential next steps**

Based on the existing limitations of research and data around exposure to pornography, there are some immediate next steps / opportunities for improving the underlying evidence base as well as the kind of data we collect that indicates how the prevalence and impact of this specific type of online harm is changing over time.

We have outlined these in the table below. Each is linked to one of the key parts of the overarching measurement process described in Section 1 [‘What could be’](#).

Which aspect of the measurement process does it relate to?	Key limitations of existing measures	Recommended next steps
<p><b>1. Input.</b> A set of assumed links between hazards, risk factors and harms.</p>	<p>There is limited understanding of the impacts of underage exposure to pornography over a long time (positive or negative)</p>	<p>More longitudinal research is needed to understand the long-term impacts and influences on people resulting from their early exposure to pornography.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Commission a new longitudinal study that explores exposure to pornography over the course of young people’s lives and the links with behaviour and attitudes over time (compared to other key factors in the development of sexual norms).</li> </ul>

Which aspect of the measurement process does it relate to?	Key limitations of existing measures	Recommended next steps
<p><b>2. Measure causes (hazards) and risk factors.</b></p> <p>Measure the prevalence of hazards and the risk factors involved.</p>	<p>Surveys are often limited to broad, imprecise definitions of pornography or content people see.</p> <p>This makes it challenging to track whether certain types of pornography are more closely linked with certain harms, or even prevalence of underage exposure.</p>	<p>Where possible, surveys should attempt to identify more precisely the types of content people have been exposed to, when, and how much.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Commission new quantitative research with the specific aim of providing a greater level of detail on experiences of exposure to pornography that can track changes over time at the population level.</li> <li>■ Provide recommendations on a question set and/or expectations for the types and granularity of data that surveys covering pornography exposure will report on (to ensure they are a valuable contribution to the overall online harms measurement process).</li> <li>■ Include a suitably granular exploration of pornography exposure in existing quantitative studies with young people.</li> </ul> <p>Note: as highlighted, there are limits to how accurate self-report methods will be when asking people about sensitive topics such as exposure to pornography. However, this data will still provide a valuable baseline, which can be tracked over time, acknowledging it is likely to be an underestimate of actual exposure.</p>
	<p>Automated tools (AI) are relatively well equipped to identify pornographic content (compared to more context-dependent hazards such as ‘abuse’) but are not utilised consistently.</p>	<p>Where there is a real intention to measure exposure to pornographic content, AI tools should be considered.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Explore the use of AI-based content detection on platforms to identify pornographic content and who (which accounts, users) are exposed to it.</li> </ul>
	<p>Platforms that hold data on users’ ages (e.g. social media) do not report on those exposed to pornographic content on their sites.</p>	<p>Where exposure to pornographic content on-site is something platforms report on, knowing the profile of those who were exposed to that content would be a valuable development.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Include this within any platform-reporting expectations/briefs.</li> </ul>

Which aspect of the measurement process does it relate to?	Key limitations of existing measures	Recommended next steps
<p><b>3. Measure impacts.</b></p> <p>Measure what actually happens to people / society who have experienced the hazards.</p>	<p>Surveys do not consistently collect data on the impact of exposure to pornography.</p>	<p>Surveys should ask about the impact of exposure to pornography, to better understand the prevalence of harm.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Develop a set of questions which ethically ask individuals about both short- and long-term impacts of inappropriate access to pornography</li> <li>■ Commission qualitative research to better understand the long-term impacts of exposure to pornography</li> </ul> <p>Note: There are limits to how much this is possible, given that some impacts will be long term and may not be recognisable to the individual.</p>

## C. Cyberbullying

### Defining cyberbullying

According to the Department of Education<sup>68</sup>:

“Bullying is usually defined as behaviour that is:

- Repeated
- Intended to hurt someone either physically or emotionally
- Often aimed at certain groups, for example because of race, religion, gender or sexual orientation

Cyber bullying is bullying via mobile phone or online (e.g. social media, instant messenger, email)”.

Several other definitions include an element of power imbalance e.g. “An aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time against a victims who cannot easily defend him or herself.” (Olweus, 1993 as cited in Smith et al., 2008)<sup>69</sup>

Some key challenges with definitions of cyberbullying include:

**Subjectivity:** what will ‘hurt’ one person will be different to what ‘hurts’ another – and may also depend on who is carrying out the bullying behaviour

**Inconsistent definitions used in this space.** For example, one academic mentioned that some cyberbullying research uses a definition which requires an ‘imbalance of power’ between the bully and victim, and others don’t, which can result in differences in estimates of the prevalence of cyberbullying – e.g. where research does not require an imbalance of power for it to be counted as ‘cyberbullying’ estimates are likely to be higher.

**Cyberbullying behaviours change over time.** The types of bullying behaviours change depending on the platforms people are using, and online ‘trends’.

### Summary of the current measures of cyberbullying

Surveys are the most commonly used way to measure the prevalence of cyberbullying. Given the subjectivity and necessary context involved in whether or not something is hurtful, it is likely that measures of cyberbullying will rely somewhat on self-report. There are several annual surveys which aim to measure the prevalence of cyberbullying over time—for example, Ditch the Label’s Bullying survey<sup>70</sup>, Ofcom’s Children and Parents: Media use and attitudes report<sup>71</sup>, and the Office for National Statistics 10–15-year-olds’ Crime Survey for England & Wales<sup>72</sup>.

Surveys use a variety of techniques to define ‘cyberbullying’. For example, Ditch the Label asks people to use their own definition of ‘cyberbullying’ to determine whether or not they feel they have experienced or witnessed it. On the other hand, the 10–15-year olds’ Crime Survey for England and Wales (CSEW) showed participants a list of ‘bullying behaviours’ and asked whether they had experienced any. 52% of those who had experienced an online bullying behaviour said they would not describe their experiences as “bullying”.

68 GOV.UK. 2021. [Bullying at school](#). [online]

69 Smith, P., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S. and Tippett, N., 2008. Cyberbullying: its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry*, 49(4), pp.376-385.

70 Ditch the Label. 2020. [The Annual Bullying Survey 2020 | Ditch the Label](#). [online]

71 Ofcom, 2020. [Children and parents: Media use and attitudes report 2019](#). Making Sense of Media. Ofcom. [online]

72 Office for National Statistics. 2020. [Online bullying in England and Wales - Office for National Statistics](#). [online]

This illustrates how subjective the term “bullying/cyberbullying” is – with differences in what researchers may consider to be bullying and what young people would say classified.

Some of these surveys also talk about the impact of bullying – for example, Ditch the Label reports various impacts such as running away from home, feeling anxious and self-harming. However, this is not specific to bullying experienced online and instead refers to the impact of all experiences of bullying, many of which occurred offline. Similarly, the Crime Survey for England and Wales (CSEW) provides an overview of whether those experiencing bullying behaviours felt ‘emotionally affected’, while Ofcom Media Use and Attitudes does not cover the impact. Understanding of the impact of cyberbullying, and how this is associated with the type and frequency of cyberbullying seems to have been given less weight.

*“Often how severe the bullying was, and how cyberbullying affected someone isn’t given as much weight [compared to how often people experienced cyberbullying behaviours]”*

#### Academic, Cyberbullying

While some surveys highlight which platforms cyberbullying took place on, and where on a platform this is likely to happen (e.g. the CSEW highlights that a large proportion of cyberbullying happens on group or private messages), this was not consistent. The CSEW also provides breakdowns for how the prevalence of online bullying differs for those with disabilities, of different ethnicities and genders.

Attempts to measure cyberbullying using AI have pointed out that more “fine-grained” or clearly defined forms of cyberbullying such as threats, expressions of racism and curses may be more likely to be identified with higher precision, when compared to cyberbullying that is less “explicit” and therefore harder for AI models to detect<sup>73</sup>.

## Key limitations and potential next steps

Based on the existing limitations of research and data around cyberbullying, there are some immediate next steps / opportunities for improving the underlying evidence base, as well as the kind of data we collect that indicates how the prevalence and impact of this specific type of online harm is changing over time.

We have outlined these in the table below. Each is linked to one of the key parts of the overarching measurement process described in [Section 1](#).

---

<sup>73</sup> Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W. and Hoste, V., 2018. Automatic detection of cyberbullying in social media text. *PLOS ONE*, 13(10), p.e0203794.

Which aspect of the measurement process does it relate to?	Key limitations of existing measures	Recommended next steps
<p><b>1. Input.</b></p> <p>A set of assumed links between hazards, risk factors and harms.</p>	<p>Few qualitative studies explore what types of cyberbullying behaviours are causing harm and how this changes over time.</p>	<p>Qualitative, longitudinal research is needed to understand the links between different types/expressions of cyberbullying and the impact this has on individuals, and in which circumstances.</p> <p>The wealth of knowledge from experts working with children around bullying (teachers, academics, support workers, charities etc.) should be used to inform central understanding of the key components of cyberbullying that are believed to cause greatest harm.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Commission new longitudinal work that explores children's exposure to cyberbullying over the course of their childhood, the links with offline bullying, and their behaviour and attitudes over time (compared to other key factors that could lead to bullying-related harm).</li> <li>■ Include an exploration of exposure to cyberbullying and the impacts of this in existing longitudinal work (e.g. existing cohort studies).</li> <li>■ Compile known links between cyberbullying experience and children's behaviour from experts. Highlight knowledge gaps where existing data/evidence may provide valuable answers.</li> </ul>

Which aspect of the measurement process does it relate to?	Key limitations of existing measures	Recommended next steps
<p><b>2. Measure causes (hazards) and risk factors.</b></p> <p>Measure the prevalence of hazards and the risk factors involved</p>	<p>Surveys do not always provide information on where cyberbullying took place – e.g. which platforms and where on the platform.</p>	<p>Surveys should gather data on where cyberbullying took place: which platforms, and which features within platforms (e.g. Private message, group message, public feed).</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Provide recommendations on a question set and/or expectations for the types and granularity of data that surveys covering cyberbullying will report on (to ensure they are a valuable contribution to the overall online harms measurement process).</li> <li>■ Include a suitably granular exploration of cyberbullying in existing quantitative studies with young people.</li> </ul> <p>Qualitative / explorative research is needed to ensure understanding of the types of behaviours which constitute ‘cyberbullying’ are up to date, and therefore can be asked in surveys, and aim to be identified by platforms/AI.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Ensure definitions for cyberbullying are updated and consistently applied, allowing different actors (researchers, platforms etc.) to collect comparable data</li> </ul>
	<p>Automated tools can pick up on language that has already been flagged/ understood to have the potential to be a hazard.</p> <p>There are inconsistencies in their ability to detect what is a hazard, many of the images in particular are unusual/unique and therefore a trained model will struggle to find those new ones.</p>	<p>Industry wide sharing of taxonomies of potentially harmful terms, hashtags, images etc.</p> <p>Technological development in AI to better detect cyberbullying.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Ensure definitions for cyberbullying are updated and consistently applied, allowing different actors (researchers, platforms etc.) to collect comparable data.</li> <li>■ Encourage and enable the sharing of data and learnings between platforms and actors attempting to identify cyberbullying.</li> </ul>

Which aspect of the measurement process does it relate to?	Key limitations of existing measures	Recommended next steps
	<p>Platforms do not report in any granular way on which users are exposed to content considered to be 'cyberbullying'.</p>	<p>Platform reporting needs to account for a wider range of key risk factors.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Include this within any platform-reporting expectations/briefs.</li> </ul>
	<p>Surveys and qualitative research do tend to identify individual characteristics (e.g. sexuality) that appear to be linked with greater likelihood of experiencing cyberbullying. However, they do not explore other key contextual factors such as the volume of cyberbullying an individual is exposed to, the social context in which people are exposed to it, the severity of the related content/actions etc.</p>	<p>Existing or new measurement needs to account for wider contextual factors wherever possible – especially where these are linked with worse outcomes/impacts (i.e. from other parts of this process).</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Provide centralised recommendations on what contextual factors need to be understood. These can inform the design of any new research/data collection.</li> <li>■ Identify gaps in knowledge and where additional risk factors need to be explored</li> </ul>



Which aspect of the measurement process does it relate to?	Key limitations of existing measures	Recommended next steps
<p><b>3. Measure impacts.</b></p> <p>Measure what actually happens to people / society who have experienced the hazards.</p>	<p>Surveys do not consistently collect data on the impact of cyberbullying.</p>	<p>Surveys should ask about the impact of cyberbullying, to better understand the prevalence of harm.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Provide centralised recommendations on the potential impacts that need to be understood/captured in any relevant surveys in order for the data collected to be a valuable addition to online harm measurement.</li> <li>■ Platforms should collect information that has been identified as being a clear indicator of harm experienced as a result of cyberbullying which may act as a flag that a user has experienced harm (e.g. sudden drop in engagement/usage).</li> </ul> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Include this information within any platform-reporting expectations/briefs.</li> </ul>

## D. The promotion of illegal / risky / dangerous behaviour

### The scope and definition of risky / illegal / dangerous behaviour

A wide array of behaviours fit into the category of risky, dangerous and/or illegal behaviours. This includes behaviour which is clearly illegal such as underage drinking and drug abuse, as well as actions that exist in a grey area such as promoting dangerous stunts online or promoting eating disorders and self-harm.

This section will focus primarily on the promotion of self-harm, suicide and eating disorders online ([Section D1](#)) as these areas are more clearly defined and understood than others and are generally considered to be particularly important 'legal but harmful' areas for online harms legislation to be tackling.

There is a separate, shorter, section on the promotion of dangerous and risky behaviour—often considered to refer to stunts and pranks—and illegal activity which includes the promotion of underage drinking, drug abuse, grooming and gangs ([Section D2](#)). As with the rest of this report, the focus on illegal activity is limited due to the range of additional laws, regulations and powers relating to these areas.

### How can behaviours be promoted online?

Before looking at the specific definitions of these different types of potential harms, it is useful to consider how these behaviours can be promoted online. The harms listed above can be all be promoted online – with 'promotion' defined as anything which **makes a behaviour easier to do or that motivates people to do it more.**

Promoting harmful behaviours online can take a number of different forms including:

- Influencing or encouraging an individual to do something
- Offering instructions and information about how to do a particular activity
- Providing knowledge about an activity
- Fostering a sense of community and engagement around a particular topic
- Making certain behaviour or actions seem aspirational or glamorous and appealing
- Making a behaviour or action seem normal

# D1. The promotion of eating disorders, self-harm and suicide

## Defining the promotion of eating disorders, self-harm and suicide

There has been increasing attention about the potential for eating disorders, self-harm and suicide to be promoted online and the potential health risks this poses. The hazards associated with these harms typically come in the form of content (posts, images, videos) and from interactions with others (comments, closed group discussions).

**Eating disorders:** The NHS defines eating disorders as “a mental health condition where you use the control of food to cope with feelings and other situations”<sup>74</sup>. Having an eating disorder is associated with a change in behaviour, thoughts and emotions which is characterised by a preoccupation with food, body weight and shape<sup>75</sup>.

Beat, the UK’s leading eating disorder charity, estimates that 1.25 million people in the UK suffer from an eating disorder<sup>76</sup>. Anorexia, one of the most common types of eating disorder, has the highest mortality rate of any mental illness<sup>75</sup>.

**Self-harm:** The mental health charity Mind describes self-harm as “when you hurt yourself as a way of dealing with very difficult feelings, painful memories or overwhelming situations and experiences”<sup>77</sup>. This is described a wide range of different behaviours and methods and therefore the ways in which it can be promoted online are diverse and also evolve and change over time.

**Suicide:** the act of intentionally taking one’s own life<sup>78</sup>. In a similar way to self-harm hazards, the promotion of suicide online can take many different forms and the degree to which something might cause harm to an individual is largely dependent on the individual who comes into contact with it.

The Samaritans’ Online Excellence Programme provides detail on content which can be harmful for users in their guidance for platforms<sup>79</sup>.

## Summary of the current measures of the promotion of eating disorders, self-harm and suicide

### Measures of the prevalence of content promoting eating disorders, self-harm and suicide

There are a number of surveys which have asked about the prevalence of this type of content. For example, Ofcom’s Internet users’ concern about and experience of potential online harms<sup>18</sup>, London School of Economics’ EU Kids Online<sup>24</sup> as well as one-off surveys such as Young people who access harm-advocating

---

74 NHS UK. 2021. [Overview – Eating disorders](#). [online].

75 NIMH. 2021. [NIMH » Eating Disorders](#). [online].

76 Beat. 2021. [About Beat - Beat](#). [online].

77 Mind. 2021. [What is self-harm?](#). [online].

78 NHS Scotland - Inform. 2021. [Suicide information](#). [online].

79 Samaritans, n.d. [Developing and implementing self-harm & suicide content policies](#). Samaritans. [online].

online content: A four-country survey<sup>80</sup>, and a cross sectional study self-report questionnaire run with participants of the Avon Longitudinal Study of Parents and Children<sup>81</sup>.

The majority of these surveys are based on questions that ask whether the individual has seen a certain type of content, either 'ever' or in a set time period such as the last 12 months. These questions rely on the individual knowing that they have seen something which falls into the categories of eating disorders, self-harm and suicide and do not provide information about the context the content was seen in or the severity of the content. In addition, surveys often do not include information about the frequency of which someone is looking at it or how much of their total time online is spent looking at that type of content.

While social media platforms report on the amount of content that they consider to promote self-harm, suicide or eating disorders on their platform, this only includes content which is detected using automated mechanisms, found by moderators or reported by users. Platforms also don't publish information about who is exposed to content (e.g. age profiles), the severity of content, or are able to understand the impact of it.

Identifying all the content that could be hazardous in this harm area is challenging. What may be harmful to one person is unlikely to be so for another. For example, lived experience accounts of self-harm may be harmful to some but seen as supportive for others. In the promotion of eating disorder content, there is a large spectrum of content that could motivate someone to develop unhealthy eating behaviours – from bikini pictures to pro anorexia pages. Work done by organisations such as the Samaritans to provide guidance on the type of content that is considered harmful<sup>79</sup> is valuable in clearly breaking down different types of self-harm and suicide content and explaining where the impacts are known and less well known.

### Measures of the impact of content promoting eating disorders, self-harm, and suicide

There are some surveys which ask users about the impact that seeing content which promotes eating disorders, self-harm and suicide had on them, as well as their exposure to content. For example, the annual Ofcom Internet users' experiences of potential online harms survey (now replaced by the Pilot Online Harms Survey) asks "what impact has [content promoting self-harm e.g. cutting, anorexia, suicide] had on you?" with a scale for answering that asks the extent to which it was "annoying, upsetting or frustrating". Clearly, this is limited in being able to understand the various ways content may have affected someone – in this instance the answer option (how "annoying, upsetting or frustrating" seeing the content was) is very broad, so that the same question can be asked about the impact of other online harms covered in the survey, such as spam.

Other research aiming to establish the link between online experiences and harm has focused on those who are known to have experienced harms relating to eating disorders, self-harm and suicide. For example, a survey by the Samaritans and University of Bristol surveyed 8,000 individuals hospitalised as a result of suicide attempts, to understand their use of the internet – with 8% of those in hospital following a suicide attempt saying they had used the internet in connection with their attempt<sup>31</sup>.

The Samaritans are currently running a number of research projects aimed at getting a better understanding of how self-harm and suicide content differs across platforms and what its impact is on users. This includes a survey with over 16s in the UK which asks about online experiences of posting or coming across self-harm and suicide content online, which aims to better understand what types of hazards (i.e. content that would be considered a 'hazard') are more likely to be harmful or helpful to people<sup>82</sup>.

Other research that charities and academics spoke about currently working on during expert interviews including running online ethnographies to explore the comments under self-harm and suicide content, and longitudinal diary studies with individuals who have attempted suicide, to understand more about their

80 Oksanen, A., Näsi, M., Minkkinen, J., Keipi, T., Kaakinen, M. and Räsänen, P., 2016. Young people who access harm-advertising online content: A four-country survey. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 10(2).

81 Mars, B., Heron, J., Biddle, L., Donovan, J., Holley, R., Piper, M., Potokar, J., Wylie, C. and Gunnell, D., 2015. Exposure to, and searching for, information about suicide and self-harm on the Internet: Prevalence and predictors in a population based cohort of young adults. *Journal of Affective Disorders*, 185, pp.239-245.

82 Samaritans. 2021. [Samaritans' online harms research launch](#). [online].

online behaviours. In particular, this would enable researchers to learn more about what happens in 'invisible' encrypted spaces such as private messages or group chats.

While qualitative studies like these are able to start to understand the complex links between hazards and harms, they are relatively small scale, expensive, and reporting may take some time. Furthermore, this type of research requires careful consideration, as well as time and money, to ensure it is carried out ethically.

It is also important to note that much of the research understanding the links between hazardous content and harm in this area has been carried out with individuals who have experienced more extreme harms i.e. have developed an eating disorder, self-harmed or attempted suicide. While this research is incredibly valuable, there are likely many people who will never experience such harms, but will nonetheless be affected by content promoting self-harm, suicide or eating disorders.

Wider datasets on the prevalence of eating disorders, self-harm and suicide already exist and are regularly updated, which offer information about how prevalent the harm is within the country as a whole. They may be helpful to monitor or combine with other data on online content. Two key datasets the government holds on these harms are: death records<sup>83</sup> and emergency hospital admissions<sup>49</sup>. The death records record when someone has died as a direct result of either an eating disorder, self-harm or suicide, while the hospital records only show who has been admitted as an emergency which can be attributed directly to one of the aforementioned harms. The NHS also runs a survey called the Mental Health of Children and Young People in England<sup>84</sup> which has run in 2004, 2017 and 2020, asking questions about mental health more generally, which includes questions about eating disorders, self-harm and suicide, but does not have an online component as part of the survey. Again, these data sets only represent a subset of harms occurring as a result of online content and are not currently linked to online experiences.

## Key limitations and potential next steps for the promotion of eating disorders, self-harm and suicide

Based on the existing limitations of research and data around eating disorders, self-harm and suicide, there are some immediate next steps / opportunities for improving the underlying evidence base, as well as the kind of data collected, that indicates how the prevalence and impact of this specific type of online harm is changing over time.

We have outlined these in the table below. Each is linked to one of the key parts of the overarching measurement process described in Section 1 '[What could be](#)'.

---

83 Office for National Statistics. 2020. [Deaths from eating disorders and other mental illnesses - Office for National Statistics](#). [online].

84 NHS Digital. 2020. [Mental Health of Children and Young People in England, 2020: Wave 1 follow up to the 2017 survey](#). [online]

Which aspect of the measurement process does it relate to?	Key limitations of existing measures	Recommended next steps
<p><b>1. Input.</b></p> <p>A set of assumed links between hazards, risk factors and harms.</p>	<p>There are some studies in this area that are working to understand the links between hazards and harms such as work by the Samaritans. However, these studies are small scale, expensive to run and require careful ethical management. Unless they are repeated continuously, they also run the risk of becoming out of date relatively quickly.</p>	<p>Investment in qualitative longitudinal research is needed to better understand the links between the content an individual comes into contact with, their offline activities, and the impact this has on their day-to-day life – understanding both the positive and negative impacts</p> <p>Establishing which hazards are the most likely to cause harm is also important so that those hazards can be prioritised for appropriate mitigation.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Commission new longitudinal work that explores children’s exposure to eating disorder, self-harm and suicide content over the course of their childhood, and their behaviour and attitudes over time (compared to other influences on their behaviour and attitudes).</li> <li>■ Include an exploration of this content in encrypted spaces.</li> <li>■ Work with individuals who understand how eating disorders, self-harm and suicide manifest offline to understand how, where and what role the online world can play in this both negatively but also positively.</li> </ul>

Which aspect of the measurement process does it relate to?	Key limitations of existing measures	Recommended next steps
<p><b>2. Measure causes (hazards) and risk factors.</b></p> <p>Measure the prevalence of hazards and the risk factors involved.</p>	<p>Surveys about mental health rarely include information or questions about the role that online activities have on someone.</p>	<p>Surveys should include information on which platforms people have engaged with eating disorder, self-harm and suicide content, and where on platform (e.g. Private message, group message, public feed), and what type of content they engaged with</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Provide recommendations on a question set and/or expectations for the granularity of data that surveys covering eating disorder, self-harm and suicide content should report on (to ensure they are a valuable contribution to the overall online harms measurement process).</li> </ul>
	<p>The type of content that promotes eating disorders, self-harm and suicide is likely to evolve and change over time, and it is challenging for measures to stay up to date.</p>	<p>A continuous mapping exercise should be undertaken to understand the content that promotes eating disorders, self-harm and suicide online. This would ensure that surveys, interviews and platform reporting remain up to date.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Work with organisations, such as the Samaritans, to develop clear and up to date guidelines on the types of content that may promote eating disorders, self-harm and suicide</li> <li>■ Ensure that information and data about the changes to the content that may promote eating disorders, self-harm and suicide are shared across platforms. Ensure this is incorporated into the development of AI models.</li> </ul>
	<p>There is a lack of granularity in how platforms report on who sees and interacts with content promoting eating disorders, self-harm and suicide.</p>	<p>Platforms should aim to report aggregate data on who is engaging with content that promotes eating disorders, self-harm or suicide, and how much – e.g. are there some users whose feeds contain a high proportion of this content</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Include this within any platform-reporting expectations/briefs.</li> </ul>

Which aspect of the measurement process does it relate to?	Key limitations of existing measures	Recommended next steps
<p><b>3. Measure impacts.</b></p> <p>Measure what actually happens to people / society who have experienced the hazards.</p>	<p>There is limited research on the impacts of exposure to this type of content, beyond those who have experienced extreme harm.</p>	<p>Surveys should ask about the impact of seeing content that promotes eating disorders, self-harm and suicide for individuals who already suffer from one of these harms, as well as for those who have not previously suffered from these harms, to see the impact on those different individuals of seeing that content.</p> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Provide centralised recommendations on the potential impacts that need to be understood/ captured in any relevant data collection (to ensure they are a valuable contribution to the overall online harms measurement process).</li> <li>■ Platforms should collect information that has been identified as being a clear indicator of harm experienced as a result of the promotion of eating disorders, self-harm and suicide which may act as a flag that a user has experienced harm (e.g. sudden drop in engagement/usage).</li> </ul> <p><b>Potential actions:</b></p> <ul style="list-style-type: none"> <li>■ Include this information within any platform-reporting expectations/briefs.</li> </ul>



## D2. Promotion of dangerous and risky behaviour, and illegal activity

### Defining the promotion of dangerous behaviour and illegal activity

As stated at the start of this section, the promotion of dangerous, risky and illegal activity represents a wide range of content.

**Dangerous behaviours** in the context of this report refer to content or interactions that promote dangerous challenges, stunts and pranks. YouTube's transparency report explains that it removes content and asks people not to post content that they consider to be:

- “extremely dangerous challenges: challenges pose an imminent risk of physical injury”
- “Dangerous or threatening pranks: Pranks that lead victims to fear imminent serious physical danger or that create serious emotional distress in minors.”

The promotion of **illegal activity** relates to behaviours such as: promoting underage drinking, drug abuse, or gang activity.

This research did not carry out in depth research into specific dangerous or illegal behaviours, but looked at examples of the types of measures used in this harm area.

### Summary of current measures of dangerous behaviour and illegal activity

#### Prevalence of content promoting dangerous/risky behaviour

Overall, there is very little research and understanding of the link between coming across or engaging with dangerous/risky behaviour and the potential for this to cause someone harm. Trending pranks or stunts that go wrong are often spoken about in the media, for example the Bird Box challenge (when someone films themselves having to navigate situations blind folded)<sup>85</sup>.

Video based platforms such as YouTube and TikTok report on the prevalence of content that they remove for violating their community guidelines around dangerous content. YouTube describes dangerous behaviour or threatening pranks as “pranks that lead victims to fear imminent serious physical danger or that create serious emotional distress on the part of minors”<sup>17</sup>.

There are some studies that discuss how dangerous activities are promoted online, as part of Ofcom's Internet users' experience of potential online harms in 2020 pranks came up as a concern for some children, however this was unprompted and there are not specific questions dedicated to pranks in the survey<sup>18</sup>.

The research into the promotion of dangerous and illegal behaviour is limited, however there are some academic studies that have looked into it. Researchers from the University of Durham ran a one-off survey looking at how social media can encourage risky behaviour. The risky behaviours asked about in the survey included: “illegal drug use, excessive alcohol consumption, extreme dieting or disordered eating, self-harm, violence on others, unprotected sex, sex with a stranger, dangerous pranks and bullying or hatred towards specific groups”<sup>86</sup>. While this study did find that there seemed to be some relationship between viewing content of ‘risky behaviour’ online

---

85 Vice. 2019. [How YouTube's Ban on Dangerous Stunts Will Affect Creators](#). [online]

86 Branley, D. and Covey, J., 2017. Is exposure to online content depicting risky behavior related to viewers' own risky behavior offline?. *Computers in Human Behavior*, 75, pp.283-287.

and offline actions, it was a small study run as a one-off and therefore more regular and large-scale testing would be needed to better understand the link between the online and offline world.

The research into the promotion of illegal activity such as the promotion of gangs is also limited. Some studies exist that use AI to try and detect gang activity. A study from the US used AI to track gang activity on social media through analysing the posts of gang members, which enabled them to understand “the structure, function and operation of the gang online”<sup>87</sup>. While these studies show how gang members use social media to promote illegal activity, it does not tell the reader anything about the prevalence of gangs promoting their activities online.

The social media hub from the Home Office’s Serious Violence Strategy will be proactively alerting social media companies of content that is promoting gang-related content online.<sup>88</sup> While this is not a measurement of the prevalence of gang-related activity online, it will add to the understanding of what the online promotion of gang activity looks like.

## Key limitations of current measures of dangerous behaviour and illegal activity

Given the research did not cover dangerous, illegal, or risky behaviours in detail beyond self-harm, eating disorders and suicide, as discussed in the previous section, it is difficult to make recommendations for next steps. However, it appears that further research is required to understand the link between content promoting dangerous or illegal behaviours, and harm. This research should seek to clarify the types of dangerous or illegal behaviour that is being promoted online.

---

87 Wijeratne, S., Doran, D., Sheth, A. and Dustin, J., 2015. Analyzing the social media footprint of street gangs. 2015 IEEE International Conference on Intelligence and Security Informatics (ISI).

88 GOV.UK. 2018. [Social media hub announced to tackle gang-related online content](#). [online].

APPENDIX 1:

**List of sources  
reviewed when  
mapping the  
measures in each  
harm area**

## Online abuse

Method	Source	Link
Evidence review	<b>Alan Turing Institute</b> How much online abuse is there? A systematic review of evidence for the UK, 2019	<a href="https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_full_24.11.2019_-_formatted_0.pdf">https://www.turing.ac.uk/sites/default/files/2019-11/online_abuse_prevalence_full_24.11.2019_-_formatted_0.pdf</a>
Evidence review	<b>University of East London, LSE, UKIS</b> (UK Council for Internet Safety) Adult Online Hate, Harassment and Abuse: A Rapid Evidence Assessment, 2019	<a href="https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/811450/Adult_Online_Harms_Report_2019.pdf">https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/811450/Adult_Online_Harms_Report_2019.pdf</a>
Survey (includes some qualitative data from Ofcom's Adults' Media lives research)	<b>Ofcom</b> Adult's Media use and Attitudes, 2020	<a href="https://www.ofcom.org.uk/__data/assets/pdf_file/0031/196375/adults-media-use-and-attitudes-2020-report.pdf">https://www.ofcom.org.uk/__data/assets/pdf_file/0031/196375/adults-media-use-and-attitudes-2020-report.pdf</a>
Survey	<b>Ofcom</b> Ofcom Pilot Online Harms survey, 2021	<a href="https://www.ofcom.org.uk/__data/assets/pdf_file/0014/220622/online-harms-survey-waves-1-4-2021.pdf">https://www.ofcom.org.uk/__data/assets/pdf_file/0014/220622/online-harms-survey-waves-1-4-2021.pdf</a>
Survey	<b>Ofcom</b> Internet users' experience of potential online harms: summary of survey research, 2020	<a href="https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/internet-use-and-attitudes/internet-users-experience-of-harm-online">https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/internet-use-and-attitudes/internet-users-experience-of-harm-online</a>
Survey	<b>Galop</b> Online Hate Crime Report: Challenging online homophobia, biphobia and transphobia, 2020	<a href="https://www.report-it.org.uk/files/online-crime-2020_0.pdf">https://www.report-it.org.uk/files/online-crime-2020_0.pdf</a>
Survey	<b>Amnesty</b> Amnesty reveals alarming impact of online abuse against women, 2017	<a href="https://www.amnesty.org/en/latest/news/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/">https://www.amnesty.org/en/latest/news/2017/11/amnesty-reveals-alarming-impact-of-online-abuse-against-women/</a>
Survey	<b>Thomas et al.</b> SoK: Hate, Harassment, and the Changing Landscape of Online Abuse, 2018	<a href="https://rist.tech.cornell.edu/papers/sok-abuse.pdf">https://rist.tech.cornell.edu/papers/sok-abuse.pdf</a>

Survey	<b>Pew Research Centre</b> The state of online harassment, 2021	<a href="https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/">https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/</a>
Survey	<b>Greenwood et al.</b> Online abuse of MPs from 2015-19, 2019	<a href="http://eprints.whiterose.ac.uk/145982/1/1904.11230v1.pdf">http://eprints.whiterose.ac.uk/145982/1/1904.11230v1.pdf</a>
Survey	<b>Stonewall / YouGov</b> LGBT in Britain, hate Crime and discrimination, 2017	<a href="https://www.stonewall.org.uk/system/files/lgbt_in_britain_hate_crime.pdf">https://www.stonewall.org.uk/system/files/lgbt_in_britain_hate_crime.pdf</a>
Survey	<b>The Fawcett Society</b> Online Abuse and Harassment Survey – Results, 2017	<a href="http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/home-affairs-commit">http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/home-affairs-commit</a>
Survey	<b>Walters et al.</b> Hate crimes against trans people: assessing emotions, behaviors and attitudes towards criminal justice agencies, 2020	<a href="http://sro.sussex.ac.uk/id/eprint/67633/4/Hate%20Crimes%20Against%20Trans%20People%20-%20final%20version%20for%20open%20access.pdf">http://sro.sussex.ac.uk/id/eprint/67633/4/Hate%20Crimes%20Against%20Trans%20People%20-%20final%20version%20for%20open%20access.pdf</a>
Survey	<b>James et al.</b> Aggressive/intrusive behaviours, harassment and stalking of members of the United Kingdom parliament: a prevalence study and cross-national comparison, 2016	<a href="https://www.researchgate.net/profile/David-James-15/publication/290475819_Aggressiveintrusive_behaviours_harassment_and_stalking_of_members_of_the_United_Kingdom_parliament_a_prevalence_study_and_cross-national_comparison/links/59e7903f458515c3630f9580/Aggressive-intrusive-behaviours-harassment-and-stalking-of-members-of-the-United-Kingdom-parliament-a-prevalence-study-and-cross-national-comparison.pdf">https://www.researchgate.net/profile/David-James-15/publication/290475819_Aggressiveintrusive_behaviours_harassment_and_stalking_of_members_of_the_United_Kingdom_parliament_a_prevalence_study_and_cross-national_comparison/links/59e7903f458515c3630f9580/Aggressive-intrusive-behaviours-harassment-and-stalking-of-members-of-the-United-Kingdom-parliament-a-prevalence-study-and-cross-national-comparison.pdf</a>
Measurement study using automated tools	<b>Farrell et al.</b> MP Twitter Engagement and Abuse Post-first COVID-19 Lockdown in the UK: White Paper, 2021	<a href="https://arxiv.org/abs/2103.02917">https://arxiv.org/abs/2103.02917</a>
Measurement study using automated tools	<b>Zannettou et al.</b> What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber? 2018	<a href="https://arxiv.org/pdf/1802.05287.pdf">https://arxiv.org/pdf/1802.05287.pdf</a>

Feasibility study of using automated tools	<b>Demos</b> Signal and Noise: Can technology provide a window into the new world of digital politics in the UK? 2017	<a href="https://www.demos.co.uk/wp-content/uploads/2017/05/Signal-and-Noise-Demos.pdf">https://www.demos.co.uk/wp-content/uploads/2017/05/Signal-and-Noise-Demos.pdf</a>
Automated tools	<b>Centre for the Analysis of Social Media (Demos &amp; the University of Sussex)</b> The use of misogynistic terms on Twitter, 2016	<a href="https://demosuk.wpengine.com/wp-content/uploads/2016/05/Misogyny-online.pdf">https://demosuk.wpengine.com/wp-content/uploads/2016/05/Misogyny-online.pdf</a>
House of Commons report	<b>House of Commons Petitions Committee</b> Online abuse and the experience of disabled people, 2019	<a href="https://publications.parliament.uk/pa/cm201719/cmselect/cmpetitions/759/759.pdf">https://publications.parliament.uk/pa/cm201719/cmselect/cmpetitions/759/759.pdf</a>
House of Commons report	<b>Committee on standards in Public Life</b> Intimidation in Public Life A Review by the Committee on Standards in Public Life, 2017	<a href="chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/666927/6.3637_CO_v6_061217_Web3.1__2_.pdf">chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/666927/6.3637_CO_v6_061217_Web3.1__2_.pdf</a>
Public reporting site	<b>Tell Mama</b> Normalising hatred, Tell MAMA Annual Report, 2018	<a href="https://tellmamauk.org/wp-content/uploads/2019/09/Tell%20MAMA%20Annual%20Report%202018%20_%20Normalising%20Hate.pdf">https://tellmamauk.org/wp-content/uploads/2019/09/Tell%20MAMA%20Annual%20Report%202018%20_%20Normalising%20Hate.pdf</a>
Public reporting site	<b>CST (Community Security Trust)</b> Hidden Hate: What Google searches tell us about antisemitism today, 2019	<a href="https://cst.org.uk/public/data/file/a/b/APT%20Google%20Report%202019.pdf">https://cst.org.uk/public/data/file/a/b/APT%20Google%20Report%202019.pdf</a>
Public / Government data set	<b>Home Office</b> Hate Crime data, 2017/18	<a href="https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/748598/hate-crime-1718-hosb2018.pdf">https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/748598/hate-crime-1718-hosb2018.pdf</a>

## Inappropriate access to content (pornography)

Method	Source	Link
Survey & Qualitative research	<b>BBFC</b> Young people, Pornography & Age-verification, 2020	<a href="https://www.revealingreality.co.uk/wp-content/uploads/2020/01/BBFC-Young-people-and-pornography-Final-report-2401.pdf">https://www.revealingreality.co.uk/wp-content/uploads/2020/01/BBFC-Young-people-and-pornography-Final-report-2401.pdf</a>
Survey	<b>Ofcom &amp; ICO</b> Internet users' concerns about and experience of potential online harms, 2019	<a href="https://www.ofcom.org.uk/__data/assets/pdf_file/0028/149068/online-harms-chart-pack.pdf">https://www.ofcom.org.uk/__data/assets/pdf_file/0028/149068/online-harms-chart-pack.pdf</a>
Survey	<b>Thurman and Obster</b> The regulation of internet pornography: What a survey of under 18s tells us about the necessity for and potential efficacy of emerging legislative approaches, 2021	<a href="https://onlinelibrary.wiley.com/doi/epdf/10.1002/poi3.250">https://onlinelibrary.wiley.com/doi/epdf/10.1002/poi3.250</a>
Public reporting site	<b>UK Safer Internet Centre</b> Reports to UK's Revenge Porn Helpline, 2020	<a href="https://www.saferinternet.org.uk/blog/revenge-porn-pandemic-rise-reports-shows-no-sign-slowing-even-lockdown-eases">https://www.saferinternet.org.uk/blog/revenge-porn-pandemic-rise-reports-shows-no-sign-slowing-even-lockdown-eases</a>
Automated tools	<b>Perez et al.</b> Video pornography detection through deep learning techniques and motion information, 2017	<a href="https://www.sciencedirect.com/science/article/abs/pii/S0925231216314928">https://www.sciencedirect.com/science/article/abs/pii/S0925231216314928</a>
Automated tools	<b>Yoti</b> Anonymous Age Estimation: A Deep Dive, 2021	<a href="https://www.yoti.com/resources/yoti-age-white-paper/">https://www.yoti.com/resources/yoti-age-white-paper/</a>

## Cyberbullying

Method	Source	Link
Survey	<b>Ditch the Label</b> The Annual Bullying Survey, 2020	<a href="https://www.ditchthelabel.org/research-papers/the-annual-bullying-survey-2020/">https://www.ditchthelabel.org/research-papers/the-annual-bullying-survey-2020/</a>
Survey (also includes some qualitative work)	<b>Ofcom</b> Children and Parents: Media use and attitudes report, 2019	<a href="https://www.ofcom.org.uk/__data/assets/pdf_file/0023/190616/children-media-use-attitudes-2019-report.pdf">https://www.ofcom.org.uk/__data/assets/pdf_file/0023/190616/children-media-use-attitudes-2019-report.pdf</a>
Survey	<b>London School of Economics</b> EU Kids Online, 2020	<a href="https://www.lse.ac.uk/media-and-communications/research/research-projects/eu-kids-online/eu-kids-online-2020">https://www.lse.ac.uk/media-and-communications/research/research-projects/eu-kids-online/eu-kids-online-2020</a>
Survey	<b>NHSE</b> Mental Health of Children and Young People in England, 2017	<a href="https://files.digital.nhs.uk/D7/BDF0AD/MHCYP%202017%20Appendix%20B%20-%20Questionnaire.pdf">https://files.digital.nhs.uk/D7/BDF0AD/MHCYP%202017%20Appendix%20B%20-%20Questionnaire.pdf</a>
Survey	<b>World Health Organisation</b> Health Behaviours in School-age Children, 2018	<a href="https://www.hbsc.org/england/national-report-2020.pdf">HBSC-England-National-Report-2020.pdf (hbscengland.org)</a>
Survey	<b>Rey et al.</b> European cyberbullying intervention project questionnaire, 2015	<a href="https://doi.org/10.1016/j.chb.2015.03.065">https://doi.org/10.1016/j.chb.2015.03.065</a>
Survey	<b>Mateu et al.</b> Cyberbullying and post-traumatic stress symptoms in UK adolescents, 2020	<a href="https://www.bmj.com/content/361/n8000/e000000">Cyberbullying and post-traumatic stress symptoms in UK adolescents   Archives of Disease in Childhood (bmj.com)</a>
Survey	<b>Department for Education</b> Bullying: Evidence from LSYPE2 wave 3, 2018	<a href="https://www.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/684442/bullying-evidence-from-LSYPE2-wave-3-2018.pdf">Bullying: Evidence from LSYPE2, wave 3 (publishing.service.gov.uk)</a>
Survey	<b>Oxfordshire schools</b> Oxfordshire schools anti-bullying survey, 2020	<a href="https://schools.oxfordshire.gov.uk/cms/schoolsnews/anti-bullying-survey-results">https://schools.oxfordshire.gov.uk/cms/schoolsnews/anti-bullying-survey-results</a>
Survey	<b>OECD</b> The OECD TALIS survey, 2018	<a href="https://www.oecd-ilibrary.org/sites/1d0bc92a-en/index.html?itemId=/content/publication/1d0bc92a-en">https://www.oecd-ilibrary.org/sites/1d0bc92a-en/index.html?itemId=/content/publication/1d0bc92a-en</a>



Method	Source	Link
Survey / Helpline	<b>NSPCC</b> Childline bullying report, 2015/16	<a href="https://learning.nspcc.org.uk/media/1204/what-children-are-telling-us-about-bullying-childline-bullying-report-2015-16.pdf">https://learning.nspcc.org.uk/media/1204/what-children-are-telling-us-about-bullying-childline-bullying-report-2015-16.pdf</a>
Survey	<b>Monks et al.</b> The emergence of cyberbullying: A survey of primary school pupils' perceptions and experiences, 2012	<a href="http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1009.8160&amp;rep=rep1&amp;type=pdf">http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1009.8160&amp;rep=rep1&amp;type=pdf</a>
Survey	<b>Youthworks</b> The Suffolk Cybersurvey 2017	<a href="#">Suffolk-Cybersurvey-2017-final-report.pdf</a>
Helpline	<b>Give us a shout website</b>	<a href="#">FAQ   Shout 85258 (giveusashout.org)</a>
Qualitative study	<b>Children's Commissioner</b> Life in 'likes', 2018	<a href="https://www.childrenscommissioner.gov.uk/wp-content/uploads/2018/01/Childrens-Commissioner-for-England-Life-in-Likes-3.pdf">https://www.childrenscommissioner.gov.uk/wp-content/uploads/2018/01/Childrens-Commissioner-for-England-Life-in-Likes-3.pdf</a>
Qualitative study	<b>Ofcom</b> Children's Media Lives Report, 2020	<a href="https://www.ofcom.org.uk/__data/assets/pdf_file/0027/217827/childrens-media-lives-year-7.pdf">https://www.ofcom.org.uk/__data/assets/pdf_file/0027/217827/childrens-media-lives-year-7.pdf</a>
Automated tools	<b>Chatzakou et al.</b> Mean Birds: Detecting Aggression and Bullying on Twitter Conference paper, 2017	<a href="https://www.researchgate.net/publication/318330507_Mean_Birds_Detecting_Aggression_and_Bullying_on_Twitter">https://www.researchgate.net/publication/318330507_Mean_Birds_Detecting_Aggression_and_Bullying_on_Twitter</a>
Automated tools	<b>Van Hee et al.</b> Automatic detection of cyberbullying in social media text, 2018	<a href="https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0203794#abstract0">https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0203794#abstract0</a>
Public/government data sets / Survey	<b>ONS</b> 10 – to 15- year-olds Crime Survey for England & Wales, 2020	<a href="https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/onlinebullyinginenglandandwales/yearendingmarch2020">https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/onlinebullyinginenglandandwales/yearendingmarch2020</a>

## Promotion of illegal / dangerous / risky behaviour

Method	Source	Link
<b>Promotion of eating disorders, self-harm and suicide content</b>		
Qualitative (Digital ethnography)	<b>Alberga et al.</b> Fitspiration and thinspiration: a comparison across three social networking sites, 2018	<a href="https://jeatdisord.biomedcentral.com/articles/10.1186/s40337-018-0227-x">https://jeatdisord.biomedcentral.com/articles/10.1186/s40337-018-0227-x</a>
Qualitative (Digital ethnography)	<b>5Rights</b> Pathways: How digital design puts children at risk, 2021	<a href="https://5rightsfoundation.com/uploads/Pathways-how-digital-design-puts-children-at-risk.pdf">https://5rightsfoundation.com/uploads/Pathways-how-digital-design-puts-children-at-risk.pdf</a>
Survey	<b>Peebles et al.</b> The association between levels of pro-eating disorder website usage, disordered eating and quality of life, 2012	<a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3510745/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3510745/</a>
Survey	<b>Machado et al.</b> The prevalence of eating disorders not otherwise specified, 2007	<a href="https://pubmed.ncbi.nlm.nih.gov/17173324/">https://pubmed.ncbi.nlm.nih.gov/17173324/</a>
Survey	<b>Oksanen et al.</b> Young people who access harm-advocating online content: A four-country survey, 2016	<a href="https://cyberpsychology.eu/article/view/6179/5909">https://cyberpsychology.eu/article/view/6179/5909</a>
Survey	<b>Arendt et al.</b> Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults, 2019	<a href="https://journals.sagepub.com/doi/full/10.1177/1461444819850106">https://journals.sagepub.com/doi/full/10.1177/1461444819850106</a>
Survey	<b>Turner &amp; Lefevre</b> Instagram use is linked to increased symptoms of orthorexia nervosa, 2017	<a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5440477/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5440477/</a>
Survey	<b>Ofcom</b> Internet users' concerns about and experience of potential online harms, 2020	<a href="https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/internet-use-and-attitudes/internet-users-experience-of-harm-online">https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/internet-use-and-attitudes/internet-users-experience-of-harm-online</a>
Survey	<b>Ofcom</b> Ofcom Pilot Online Harms survey, 2021	<a href="https://www.ofcom.org.uk/__data/assets/pdf_file/0014/220622/online-harms-survey-waves-1-4-2021.pdf">https://www.ofcom.org.uk/__data/assets/pdf_file/0014/220622/online-harms-survey-waves-1-4-2021.pdf</a>

Method	Source	Link
Survey	<b>Smahel et al.</b> EU Kids Online 2020: Survey results from 19 countries, 2020	<a href="https://www.lse.ac.uk/media-and-communications/research/research-projects/eu-kids-online/eu-kids-online-2020">https://www.lse.ac.uk/media-and-communications/research/research-projects/eu-kids-online/eu-kids-online-2020</a>
Survey	<b>Mars et al.</b> Exposure to, and searching for, information about suicide and self-harm on the Internet: Prevalence and predictors in a population based cohort of young adults, 2015  Young adults were part of the Avon Longitudinal Study of Parents and Children (ALSPAC)	<a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4550475/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4550475/</a>
Survey (in progress)	<b>Samaritans</b> Samaritans launch new research to understand the effects of online self-harm and suicide content and how we can make the Internet a safer space for everyone.	<a href="https://www.samaritans.org/news/samaritans-online-harms-research-launch/">https://www.samaritans.org/news/samaritans-online-harms-research-launch/</a>
Survey and qualitative interviews	<b>Biddle et al.</b> Priorities for suicide prevention: balancing the risks and opportunities of internet use, 2016	<a href="https://www.bristol.ac.uk/media-library/sites/policybristol/briefings-and-reports-pdfs/pre-2017-briefings--reports-pdfs/PolicyBristol_Report_7_2016_suicide_and_internet.pdf">https://www.bristol.ac.uk/media-library/sites/policybristol/briefings-and-reports-pdfs/pre-2017-briefings--reports-pdfs/PolicyBristol_Report_7_2016_suicide_and_internet.pdf</a>
Public / government data sets	<b>ONS</b> Deaths from eating disorders and other mental illnesses, 2020	<a href="https://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/deathsfromeatingdisordersandothermentalillnesses">https://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/deathsfromeatingdisordersandothermentalillnesses</a>
Public / government data sets	<b>Public Health England</b> Public Health Profiles: Self-harm	<a href="https://fingertips.phe.org.uk/search/self%20harm">https://fingertips.phe.org.uk/search/self%20harm</a>
Public / government data sets	<b>NHS</b> Mental Health of Children and Young People in England, 2020: Wave 1 follow up to the 2017 survey, 2020	<a href="https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2020-wave-1-follow-up">https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2020-wave-1-follow-up</a>

Method	Source	Link
Automated tools	<b>Scherr et al.</b> Detecting Intentional Self-Harm on Instagram: Development, Testing, and Validation of an Automatic Image-Recognition Algorithm to Discover Cutting-Related Posts, 2019	<a href="https://journals.sagepub.com/doi/10.1177/0894439319836389">https://journals.sagepub.com/doi/10.1177/0894439319836389</a>
<b>Promotion of other dangerous / risky / illegal behaviour</b>		
Survey	<b>Ofcom</b> Internet users' concerns about and experience of potential online harms, 2020	<a href="https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/internet-use-and-attitudes/internet-users-experience-of-harm-online">https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/internet-use-and-attitudes/internet-users-experience-of-harm-online</a>
Survey	<b>Branley &amp; Covey</b> Is exposure to online content depicting risky behavior related to viewers' own risky behavior offline? 2017	<a href="https://www.sciencedirect.com/science/article/pii/S0747563217303357">https://www.sciencedirect.com/science/article/pii/S0747563217303357</a>
Public reporting site	<b>UK Safer Internet Centre / SWGfL</b> (South West Grid for Learning) - Reporting Harmful Content	<a href="https://reportharmfulcontent.com/?lang=en">https://reportharmfulcontent.com/?lang=en</a>
Automated tools	<b>Chang et al.</b> Detecting Gang-Involved Escalation on Social Media Using Context, 2018	<a href="https://aclanthology.org/D18-1005.pdf">https://aclanthology.org/D18-1005.pdf</a>
Automated tools	<b>Wijeratne et al.</b> Analyzing the social media footprint of street gangs, 2015	<a href="https://www.researchgate.net/profile/Sanjaya-Wijeratne/publication/307738262_Analyzing_the_social_media_footprint_of_street_gangs/links/598c71a00f7e9b07d225ba41/Analyzing-the-social-media-footprint-of-street-gangs.pdf">https://www.researchgate.net/profile/Sanjaya-Wijeratne/publication/307738262_Analyzing_the_social_media_footprint_of_street_gangs/links/598c71a00f7e9b07d225ba41/Analyzing-the-social-media-footprint-of-street-gangs.pdf</a>