# Online safety: our research agenda

Published 15 April 2024

Welsh version available

# Contents

## Our areas of interest for future research

# Foreword

Research underpins everything we do at Ofcom. We can only deliver a safer life online for people in the UK, and ensure online services effectively protect users and their rights, if we hold ourselves to a standard of excellence in our research and data activity. We've invested in an extensive programme of research to date, but we are also committed to working alongside the global community of researchers with whom we share common interests.

Ofcom has always valued its relationship with external researchers. By publishing this research agenda, we hope to accelerate our engagement and collaboration around these common goals, as we continue to keep pace with the rate of change and maximise our innovation in the online landscape.

**Gill Whitehead, Ofcom's Group Director for Online Safety**

# Introduction

## Who we are

Ofcom is the UK's communications regulator. We regulate the TV, radio and video on demand sectors, fixed line telecoms, mobiles, postal services, plus the airwaves over which wireless devices operate. In 2020, we became the regulator for video-sharing platforms (VSP). As of November 2023, Ofcom is also the regulator for online safety in the UK. Our job, under the Online Safety Act, is to make sure online services protect users in the UK, and we also have statutory duties to promote media literacy. We have in-house research and data specialists developing our online safety research programme to support our regulatory duties.

## Why we've published our research agenda

We seek to use the best possible evidence to inform and underpin our online safety policy work. This requires us to leverage the wide community of academics and researchers in the UK and around the world, so that we can: strengthen our understanding of people's online lives across the many services we regulate; make best use of research on how online experiences can be improved as technology advances and harms evolve; improve awareness of our areas of interest, as well as opportunities for support, among the research community; and share the challenges of researching online safety and the methodologies available. By publishing this research agenda, we aim to:

- **be transparent** about our research interests and goals, and the challenges we face in pursuing them;
- **invite and incentivise engagement and collaboration** with the wider research community, to further shared research goals and promote innovation; and
- **demonstrate our commitment to evidence-based regulation**, and to improving and iterating our understanding of relevant issues in the long-term.

Note that this research agenda differs from a call for evidence or consultation, in that we are not seeking evidence to support current or outstanding policy proposals and products. Rather, we are seeking to highlight areas of research interest that will support our understanding in the longer term.

# How to work with us

We currently work with academic organisations in a range of ways, such as through letters of support for proposals for project funding or co-sponsoring PhD studentships.

There is no Ofcom funding linked specifically to this research agenda. However, we may be able to support your applications for academic funding grants through our letters of support and you may be able to evidence the impact your research has by its use in informing Ofcom policy development.

On occasion, Ofcom may choose to fund research in specific areas, either via an invitation to tender, or in response to an external funding call. This will be advertised via the Ofcom website, and through Contracts Finder.

Please get in touch with Ofcom at academic.engagement@ofcom.org.uk or complete the initial expression of interest form if you would like further information about our research interests and/or to discuss ways of working with us. We will ensure that you receive a response.

# Our research programme

To prepare for our online safety role, we have invested in a comprehensive and multidisciplinary programme of research, providing the evidence base that underpins our regulation.

a) Our Research & Intelligence team leads on understanding user experiences and attitudes through quantitative and qualitative consumer research and passive monitoring tools, as well as understanding the online industry through the analysis of industry datasets.

b) Our Behavioural Insights team houses our behavioural economics and psychology expertise, particularly in using online randomised controlled trials (RCTs).

c) Our Media Literacy research team publishes annual reports on adults' and children's media literacy, as well as deep dive analysis on relevant topics.

d) Our Economics & Analytics team helps us to understand the characteristics of online businesses and the incentives they have to behave in certain ways, alongside evaluating the impact of services' safety measures.

e) Our Trust & Safety, Technology Policy, and Data Innovation teams provide technical expertise in assessing current and emerging technologies used by online services.

## What we've achieved so far

We have a strong programme of bespoke research in online safety and media literacy, which we publish regularly on the Ofcom website.

### User activity and behaviours

Our surveys provide the foundation for understanding UK online behaviours from both quantitative and qualitative perspectives. To complement our survey data, we also use passive monitoring to tell us which online sites and apps people actually visit. We experimented with a children's passive monitoring pilot to test this method with child participants, too, using the data in our annual Online Nation report. We have also used online randomised control trials in the context of Behavioural Insights studies to explore how various behavioural techniques (such as nudges and boosts) can influence user behaviour in the context of online safety measures. We have explored how a serious game could be used to improve children's understanding of social media etiquette and help them to navigate their online spaces.

### Online risk and harm

Our Online Experiences Tracker allows us to monitor the level of exposure to potential harms online for UK internet users. We have also commissioned several pieces of bespoke research to explore online harms in more depth, for example: How people are harmed online; Research into risk factors that may lead children to harm online; detailed reports on online terrorism, violence and hate; Sale of prohibited goods on search services; Understanding experiences of minority beliefs online, and Understanding online communications among children. To find new ways to further understand the prevalence and nature of online harms, we have assessed the feasibility of using the avatar methodology which may inspire future research.

## Service design and characteristics

We have been building knowledge of how the functionalities and design of a service can be amended to improve user safety. Our Video-Sharing Platforms (VSP) Tracker, set up as part of our VSP regulation, provides detailed insight into this, as well as bespoke research with the BBFC on functionality of online pornography services. We have been developing our toolset to identify the services in-scope of the online safety regulation and assess risk of harm using data at scale. Several Behavioural Insights trials are underway to explore methods of increasing users' understanding of and engagement with platforms' user choice features and Terms of Service.

## Safety measures and technologies

We launched our Online Safety Lab as a secure space for internal research and development, providing an environment for conducting artificial intelligence (AI) and/or machine learning (ML) research. We've strengthened our technical understanding of the benefits and limitations of safety measures such as content moderation (Content moderation in user-to-user online services) and furthered our understanding of the role of automated content classifiers within content moderation workflows. Our ASPARC model and Interactive Services Model have helped us examine the user journeys and trust and safety measures within different online environments. Meanwhile generative AI is a continuing research focus for us, as is our work on evaluating recommender systems, particularly in evaluating algorithmic assessment methods. We publish and discuss some of our work assessing the effectiveness of safety measures through our Economics insights and discussion papers. For example, our paper on Understanding the impact of video-sharing platform (VSP) design on user behaviour explores how the design of alert messages and content-reporting mechanisms affects the decisions people make about viewing potentially harmful content and reporting such content. Our work on Hate speech classification highlights the importance of assessing the accuracy of hate speech classification by conducting our own analysis of the accuracy of commonly used hate speech classifiers.

# What we're planning to do next

Given the depth and breadth of the Online Safety Act, and the complex and innovative nature of the services in scope, we need to invest in research on an ongoing basis to ensure we broaden our evidence base and keep up to date with any relevant changes. These changes could include changes to the online services in scope of our regulation, the functionalities and risks they pose, and the potential ways in which these risks can be mitigated in effective and proportionate ways.

We will continue to experiment with innovative methods to ensure we are using a broad range of techniques to expand and strengthen our evidence base. We will prioritise investing in our programme of children's research, using a range of tools to provide a more holistic view of children's experiences, attitudes, and behaviours. This may include expanding some of our existing surveys, developing expertise in avatar methods and scaling up our children's passive measurement techniques. We will invest further in a range of datasets that can broaden our evidence base. We plan to conduct new business surveys to understand what businesses are doing to foster a safer online environment and the challenges they face in achieving that. We will continue building our evidence base on the effectiveness of services' safety measures (e.g. user facing content controls, Terms of Service etc.) and any potential unintended consequences, using a mix of behavioural trials, machine learning and quasi-experimental techniques. In the technology space, we intend to scale up our Online Safety Lab to further evaluate safety technologies and invest in generative AI research.

In addition to these, we also hope to further our knowledge in the more specific areas of interest set out below, in partnership with the wider research community.

**The rest of this document**

Our areas of interest for future research are arranged into four themes:

1. Understanding user activity and behaviour
2. Understanding online risk and harm
3. Understanding service design and characteristics
4. Understanding safety measures and technologies

For each theme we provide an overview of why this area is important to us, followed by our specific areas of interest for future research, arranged by sub-themes.

Note that there is some overlap across the four themes. Much of our own research cuts across more than one theme, and we expect this is true of external research projects and programmes too.

The themes and sub-themes in this agenda are not exhaustive of our research interests, which are broad and continually evolving due to the scale of services and diversity of content in scope.

# Theme 1: Understanding user activity and behaviour

## Overview

1. At Ofcom we prioritise understanding what people do online and their attitudes and experiences of being online. Learning more about how a person's characteristics such as their age and gender impacts their online experiences and behaviour can help us take policy decisions that make life safer online for people in the UK.

## Areas of interest for future research

### Children's online experiences

2. Ensuring that children in the UK can live a safer life online is core to the Online Safety Act so is a focus for our research too. While we already invest in lots of research in this space, it is important that our evidence base stays up to date. The considerations for us when conducting research with children include maintaining high standards of safeguarding and ethical practice while being able to accurately assess what children do online.

3. Areas of particular interest include:

   - Methodologies for understanding what content children are being exposed to online, where, and how frequently.
   - Ways we can measure the cumulative impact of harmful content on children and their reaction/responses to it.
   - Children's interaction within services designed for children's use only ("walled gardens") and within private spaces such as group chats.
   - Methodologies for understanding the relationship between online activity and children's wellbeing, including repeated exposure to harmful content.

### Vulnerable users' online experiences

4. Developing our understanding of the needs and experiences of more vulnerable people online can help us better address their user safety needs going forward. We are particularly interested in how certain user characteristics may indicate increased vulnerability online and how approaches to user safety can best reflect this.

5. Areas of particular interest include:

   - User characteristics (e.g. neurodiversity, language barriers etc.) that may make a person more vulnerable online, and the online spaces where such users are particularly vulnerable.
   - How the effectiveness of safety measures can be tailored to serve the needs of different groups of vulnerable users.

## Behavioural insights

6. Behavioural insights help us understand how consumers and businesses behave, and how people make decisions. We use these insights to inform our policy-making, improve services, and ultimately deliver better outcomes for users and citizens. Our interest lies in the interaction between how services are designed, how users behave, and how harm manifests online, as well as what drives and influences the behaviours of the businesses we regulate.

7. Areas of particular interest include:

- Factors which shape adoption, among different demographics, of emerging safety technologies.
- Design features that can be effective in increasing informed choice or empowering users to shape their online experiences.
- Approaches to evidencing the medium-to-long term impact of design features on user behaviour (for example, the effect of repeat exposure to alert warnings or repeated prompts to update content controls).
- Design features and preventative interventions that affect more complex behaviours (for example, high-risk contact that moves across platforms or risky browsing behaviour across multiple platforms).

# Theme 2: Understanding online risk and harm

## Overview

1. Understanding the nature, causes and impacts of online harm is central to our online safety duties. The Online Safety Act distinguishes between illegal content, such as child sexual abuse material, terrorism and hate content, and content which is not illegal but may be harmful to children, such as pornography and content that encourages or promotes an eating disorder. While we already have a lot of expertise and evidence about these harms, ensuring that our evidence base is up to date to reflect the latest developments will be a continuous process.

## Areas of interest for future research

### Hate and terror

2. The use of online services to incite and radicalise vulnerable people, including children, towards hate and violence poses a major risk. It can have severe and far-reaching consequences, including for targeted minorities and protected groups. In this ever-changing space, continuing to build upon our understanding of these harms across the huge range of services in scope is vital to us.

3. Areas of particular interest include:

   - Future safety measures that could be effective at mitigating against the uploading and spreading of hate and terrorist content/activity online.
   - Techniques for learning more about the behaviours and characteristics associated with perpetrators of hate speech and terrorist content/activity online.
   - Techniques for learning more about the relationship between gaming services and hate speech and extremism.

### Mis/disinformation

4. Misinformation is one of the most prevalent potential harms encountered by both adults and children online. Ofcom's duty to promote media literacy includes helping the public understand the nature and impact of mis- and disinformation, and how they can reduce their exposure to it. We have a longstanding duty to promote and research media literacy more broadly.

5. Areas of particular interest include:

   - Emerging tactics, techniques and procedures for disinformation campaigns, and any emerging actors.
   - The prevalence of mis/disinformation and its correlation with significant moments/events, such as political developments or humanitarian crises.
   - The means by which online locations become associated with mis/disinformation over time.

## Fraud

6. Fraud is the most commonly experienced and frequently reported crime in the UK, with victims of fraud often experiencing both financial loss and a negative impact on their mental health. We know that fraudsters rapidly adapt to exploit new technologies, so it is important that we improve and update our understanding continuously, too.

7. Areas of particular interest include:

- Differences in the use of online advertisements in fraudulent activity between user generated content and paid for advertising (including on search services).
- Emerging methodologies perpetrators use to coerce users into fraud.

## Violence against women and girls

8. Women and girls experience disproportionate and distinct forms of harms online. This includes a wide range of complex and interrelated harms that seek to threaten, monitor, silence and humiliate women and girls. To meet our duties of protecting users' safety and rights, it is important to continue to build our understanding of how these harms manifest, change and adapt to technological advances.

9. Areas of particular interest include:

- How gender-based abuse manifests online and how it might be affected by individual vulnerabilities (e.g. protected characteristics, public figures) or emerging technologies.
- Potential mitigations for preventing and responding to online-gendered abuse (e.g. deterrence, safety tools) and what challenges could be faced when implementing the mitigations.

## Child sexual abuse and exploitation (CSEA)

10. The sexual exploitation and abuse of children online is a persistent and growing threat, with devastating consequences for those affected. New risks are emerging as the way we interact online evolves, and we will continue to work collaboratively with stakeholders to strengthen our evidence base and make the biggest possible impact on the safety of children online.

11. Areas of particular interest include:

- Understanding the online harms landscape of CSEA (specifically online CSE, cross-platform offending, self-generated indecent imagery, gaming and emerging threats such as extended reality and generative AI).
- Understanding impact of online CSEA on victims and survivors, including impact of emerging harms.
- Evaluating the effectiveness and usability of emerging tools of moderation for online CSEA, such as automatic content classifiers based on machine learning (ML) algorithms.

## Content that is harmful to children

12. The Online Safety Act sets out certain types of content that is harmful to children. While we have developed a strong evidence base on the nature, prevalence and impact of this content, children’s online experiences and the content they encounter continues to evolve. Sections 61 and 62 of the Act list content that is harmful to children.

13. Areas of particular interest include:

- Measuring the impacts of harm from content that is harmful to children as defined in the Online Safety Act.
- Identifying additional new and emerging harms that may not be illegal but could still be harmful to children.

# Theme 3: Understanding service design and characteristics

## Overview

1. It is important that we keep our understanding of online services up to date to ensure we can identify emerging functionalities and mitigate their unintended consequences. We must be aware of the monetisation models that impact a service's design and consider how these models could affect a user's interaction with harmful content and other users. Maintaining this research will allow us to monitor emerging in-scope services and analyse how their characteristics could have an impact on the user.

## Areas of interest for future research

### How online services are designed and function

2. Understanding the characteristics and functionalities of services, as well as how they develop over time, is central to fulfilling a range of our regulatory duties. We are interested in emerging types of service, new design features that have the potential to change or influence user experiences, and any other service characteristics relevant to online safety.

3. Areas of particular interest include:

   - The implications of new types of services, such as decentralised and immersive technology services, for user safety and media literacy.
   - The techniques available to learn more about the relationship between a service's functionalities and the risk of harm to its users.
   - How different approaches to algorithmic design affect user experience, such as exposure to and engagement with certain kinds of content.

### How online services providers' business models work

4. Business models hold an important influence on how a service develops over time. We need to keep informed on how these business models operate throughout different industries to allow us to understand their motivations and anticipate emerging risks.

5. Areas of particular interest include:

   - Techniques to better understand the relationship between a service provider's business model and the risk of harm to its users.
   - How SMEs use social media and search platforms for commercial or revenue-generating purposes.
   - How services' monetisation policies can impact on user safety, and other drivers for investment in user safety.

# Theme 4: Understanding safety measures and technologies

## Overview

1. Ofcom currently conducts research to develop our skills and understanding of trust and safety measures and technologies, and it is important we keep up with the rapid rate of change and innovation in this space. It is important for us to continue learning how design of safety measures on services can impact their effectiveness at keeping users safe online.

## Areas of interest for future research

### Evaluating safety measures

2. Assessing whether services' safety measures are effective at reducing the risk of harm to UK users is an important part of the Online Safety regime. Evaluating safety measures will also help us to understand whether these create risks of unintended effects – positive or negative. This may include assessing the impact of safety interventions on competition and innovation, freedom of expression, privacy, and users' experiences. We are interested in identifying the right metrics and analytical techniques to assess the impact of different types of safety measures, where possible at scale. We also want to explore how these approaches may need to vary according to type of harm or services studied.

3. Areas of particular interest include:

   - New or emerging analytical techniques and metrics to support the evaluation of platforms' user-facing safety measures (e.g. reporting and flagging tools, user empowerment tools).
   - Whether and how safety measures can have unintended effects on user experience, user rights, and on innovation and competition – and how such effects can be measured.
   - The potential for interventions on the largest services to result in the displacement of harms and users to other and/or smaller services.
   - Analytical approaches to assessing at scale the risk of harm to UK users on online services.

### Evaluating safety tech

4. While the arrival of new technologies comes with many opportunities and benefits, it also comes with the potential for new or different harms. We need to continually evaluate the impact of new technology and safety tech measures and have the right expertise to recommend new measures in the future. We would value efforts from the wider research community to develop new approaches and methodologies that can help us in our task to assess technologies and safety tech measures.

5. Areas of particular interest include:

- The development of novel methodologies to improve safety and/or and assess the effectiveness of new safety tech measures, when evaluating multi-layered architectures as a whole or in the following areas:
  - Recommender systems
  - Age assurance
  - Privacy enhancing technology
  - Automated Content Moderation
  - Generative AI
  - ‘Deep fakes’ and synthetic content technology
- Techniques (including methods, principles and technical metrics) to ethically create and share training data that includes harmful content.

## Generative AI

6. As generative artificial intelligence systems become more sophisticated and adoption of such applications increases, we need our evidence base to keep pace. We must continue to develop our understanding of potential impacts of generative AI on different demographics of users and different types of online harms.

7. Areas of particular interest include:

- Techniques to examine the impact of generative AI on different types of harmful content and different groups of users.
- The particular impact generative AI may have on children’s online experiences.
- Techniques to ensure the ethical governance of AI tools and training datasets, for example, mitigation of bias.

## Parental controls

8. Our research indicates that parents use a range of methods to engage with their children’s online activity, and that parents’/carers’ and children’s attitudes to these methods are influenced by several factors. Understanding these factors, as well as the effectiveness of parental control tools *in practice*, is important to providing both parents and children with the support they need online.

9. Areas of particular interest include:

- The factors that influence attitudes and adherence to parental controls among parents and children.
- How parental controls operate alongside other safety measures provided by platforms.
- How to evaluate the effectiveness and any unintended consequences of parental controls.