Evaluating online safety measures

Economics Discussion Paper Series Issue 10

Published 24 May 2024

Ofcom discussion paper series in communications regulation

The discussion paper series

Ofcom is committed to encouraging debate on all aspects of media and communications regulation and to creating rigorous evidence to support its decision-making. One of the ways we do this is through publishing a series of discussion papers, extending across economics and other disciplines. The research aims to make substantial contributions to our knowledge and to generate a wider debate on the themes covered.

Disclaimer

Discussion papers contribute to the work of Ofcom by providing rigorous research and encouraging debate in areas of Ofcom's remit. Discussion papers are one source that Ofcom may refer to, and use to inform its views, in discharging its statutory functions. However, they do not necessarily represent the concluded position of Ofcom on particular matters.

In this paper we are aiming to map out good practice for evaluation of a service's safety measures using an example of a measure that some services already have in place. In setting out evaluation methods and applying this process to an example we are not suggesting that these methods are Ofcom's finalised view or that services/Ofcom will have to use these methods.

Contents

Section

1. Overview

1.1 Ofcom is the United Kingdom's (UK) communications regulator, overseeing sectors including fixed line and mobile telecoms, the airwaves on which wireless devices operate, post and TV and radio broadcasting. We have regulated video-sharing platforms (VSPs) since November 2020 and we were formally appointed as the online safety regulator in October 2023.

1.2 The new Online Safety Act (OSA) and existing VSP regulation place duties on relevant online services to protect their users from illegal content and children from certain harmful content. While there are some differences between the rules and scope of the OSA and the VSP regulation, both regulations require services to take appropriate and proportionate online safety measures reflecting the size, nature, and risks associated with their service.

1.3 Assessing and evaluating the impact of safety measures is essential to understanding whether they effectively mitigate online harms and whether there are any unintended effects on key user rights such as freedom of expression or privacy.

1.4 In this paper, we set out how a widely used evaluation framework could be applied to assess the impact and effectiveness of online safety measures. We also provide an illustrative example of what such an evaluation might look like in practice. We hope this paper will generate a discussion on the frameworks for evaluating online safety measures amongst services, academics, civil society organisations and the broader trust and safety and online safety expert community.

Summary

This paper has two parts:

PART I – An evaluation framework to measure the effectiveness of online safety measures

We discuss the importance of proportionate evaluation and present a framework that may be used to assess the effectiveness of safety measures, including how a theory of change approach can help assess the impact of a measure. We discuss different types of metrics and measurement techniques, and then highlight some of the caveats and challenges to bear in mind while carrying out the evaluation process.

PART II – Applying our approach: Evaluating the effectiveness of parental content controls

We provide a worked example that applies the evaluation approach described in Part I. Parental content controls are one of the protection measures that the VSP regulation recommends that platforms should consider. We set out a theory of change to explain causal links from inputs to intended outcomes and identify a set of candidate metrics that could be used to assess if and to what extent parental content controls can help prevent harm (taking into account any unintended outcomes).

Concluding remarks and next steps

We invite views and explain how we intend to take this work forward.

2. PART I – An evaluation framework to measure the effectiveness of online safety measures

Introduction

2.1 The OSA places duties on relevant online services to have appropriate measures to keep users safe from illegal content and, for children, from certain harmful content. Services must conduct risk assessments and children's access assessments, and ensure that they put in place appropriate and proportionate safety measures to mitigate and manage those risks, taking into account the size, nature and risks to users of their service.

2.2 Separately, VSPs established in the UK are regulated by Part 4B of the Communications Act 2003. This regulation is more limited in scope than the OSA, but like the OSA it requires VSPs to take appropriate measures to protect users, and particularly children, from certain types of harmful material.[1] Whether a measure is appropriate for either of these purposes must be determined by whether it is practicable and proportionate for it to be taken, taking into account factors which include the size and nature of the VSP's service, the profile of their users and the nature of the material available. Where a measure is taken, it must be implemented so as to achieve the purpose for which the measure is appropriate (e.g. if the measure is to protect under 18s from being able to access adult content, its implementation must achieve that purpose).

2.3 While services are not expected to be able to completely eradicate harmful or illegal content, it is essential under both UK online safety regimes that services ensure that their measures are effective in terms of protecting users from illegal or harmful content, but that they do not have unintended impacts on user rights such as privacy or freedom of expression.

2.4 Evaluation of a service's safety measures is therefore critical to understand the impact, appropriateness and efficacy of a safety measure.

2.5 The purpose of this paper is to highlight the importance of evaluation in the online safety landscape, set out a step-by-step approach that could be used to evaluate online safety measures, and flag key caveats and challenges when measuring and reporting metrics.

---

[1] All users must be protected from **relevant harmful material** which refers to any material likely to incite violence or hatred against a group of persons or a member of a group of person based on particular grounds. It also refers to the material the inclusion of which would be a criminal offence under laws relating to terrorism, child sexual abuse material, racism and xenophobia. Children must be protected from **restricted material**, which refers to videos which have or would be likely to be given an R18 certificate by the British Board of Film Classification or which have been or would like to be refused a certificate. It also includes other material that might impair the physical, mental or moral development of under-18s.[2] See section 2.5 and pages 11-12, HM Treasury, 2022.The Green book- Central Government Guidance on Appraisal and Evaluation.

2.6 We welcome discussion about the contents of this paper, different approaches to evaluation and broader considerations that may not be covered here.

Evaluation is a key pillar of online safety regulation

2.7 Evaluation is a commonly used approach for the systematic assessment of the design, implementation and outcomes of an intervention, such as a policy or regulation. Evaluation tells us whether an intervention is achieving or has achieved its objectives; how effective it is at achieving its objectives; and its overall impact, including unintended consequences. These lessons can inform any changes to the intervention, as well as future policies or regulation, and demonstrate best practice.[2]

2.8 This paper looks at evaluation in the context of online safety, and how this systematic assessment may be applied to online safety measures and initiatives.

2.9 Evaluation of online safety initiatives is critical to:

a) *Assess the effectiveness of measures that services have implemented, including detecting and measuring both*:

i) direct intended effects that lead to a safer online experience for users as a result of better detection, lower prevalence of, and exposure to, harmful content; and

ii) indirect effects and any unintended consequences, both negative (e.g. impact on users' freedom of expression) and/or positive effects, including innovation (e.g. a critical mass of services rolling out a new safety measure enhancing investment and innovation in new technologies) or effects on creators of content (e.g. in terms of the nature of content they attempt to upload to a service).

b) *Highlight best practice:* Given the novelty of OS legislation globally, we need to build knowledge in terms of how effective different safety measures are at keeping users safe. Highlighting good practice can assist services in their own choices over safety measures. It also provides evidence to inform future regulation and policy development; this includes amending guidance if safety measures are not as effective as anticipated.

The evaluation will need to occur at different stages of implementation

2.10 The process of evaluation can occur at different stages of implementation of a safety measure, including both 'ex-post' (after implementation) and 'lifecycle' (before and during implementation) evaluations.

2.11 As the impacts of a safety measure could take time to be realised, this might point to evaluation only occurring once the safety measure has been in place for a period of time. These ex-post evaluations have the benefit of allowing services and Ofcom to gather retrospective insights once a longer series of data is available and outcomes have had time to emerge. Learnings about the extent to which a safety measure achieved its intended goal can then be taken forward in the future.

[2] See section 2.5 and pages 11-12, HM Treasury, 2022.The Green book- Central Government Guidance on Appraisal and Evaluation. //assets.publishing.service.gov.uk/media/623d99f5e90e075f14254676/Green_Book_2022.pdf [Accessed 31 January 2024].

2.12 In practice, evaluation is likely to be most beneficial when it is also considered over the lifecycle of a safety measure's implementation. Before a safety measure is first designed and launched, it can be helpful to draw on lessons from any evaluations of previous or similar safety measures. This can then help inform how and why a measure is expected to work. During the process of implementation, ongoing tracking and monitoring can provide earlier indications to services that outcomes are moving in the right direction. This can also help services act promptly when things are not working as intended.

2.13 Whatever the evaluation approach adopted, it is important to identify the monitoring and evaluation criteria in advance.[3] This will help to ensure appropriate data is collected over a sufficient time period to support the evaluation.

Good practice evaluation of safety measures

2.14 Neither the OSA nor the VSP regulation mandate a specific approach to evaluating safety measures. There are likely to be different approaches to achieving this in practice.

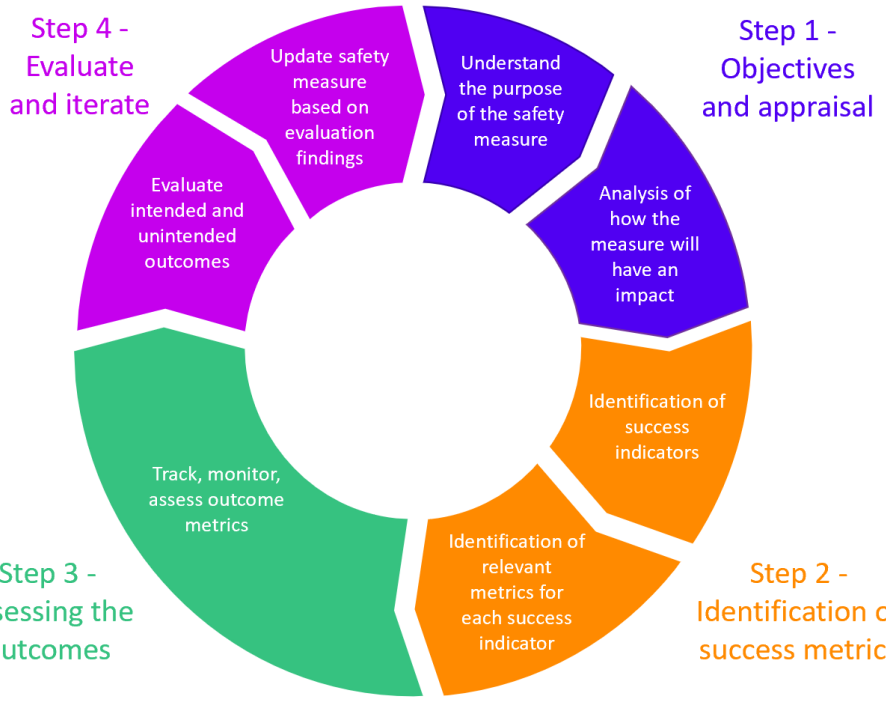
2.15 Below we take a commonly used evaluation lifecycle framework[4] and demonstrate how it could be applied to measuring the effectiveness of services' online safety measures. Following this, we discuss challenges associated with the evaluation of online safety measures. It is important to bear these challenges in mind when conducting an evaluation and particularly around what metrics and measurement approaches can best be used to assess effectiveness.

2.16 A typical evaluation lifecycle has four key steps. These steps are set out in Figure 1 and discussed in turn below.

[3] Identification of monitoring and evaluation criteria is standard practice for example where we conduct impact assessments for relevant regulatory proposals see Ofcom, 19 July 2023. Impact assessment guidance. https://www.ofcom.org.uk/__data/assets/pdf_file/0026/264707/Impact-assessment-guidance.pdf [Accessed 31 January 2024].

[4] This draws on the evaluation process presented in government guidance – see Section 1.12, page 18-20 - HM Treasury, 2020, Magenta Book – Central Government Guidance on Evaluation https://assets.publishing.service.gov.uk/media/5e96cab9d3bf7f412b2264b1/HMT_Magenta_Book.pdf

Figure 1: Evaluation lifecycle of a safety measure



Source: Ofcom

Step 1: Objectives and Appraisal

2.17 The initial step has two main stages: (a) understand the purpose of a safety measure; and (b) analysis of how the measure is likely to have an impact.

Step 1a: Understand the purpose of the safety measure.

2.18 It is important to clearly articulate the issue or problem the safety measure is designed to address and, therefore, the intended impact of the measure in improving the safety of the service's users.

2.19 The safety measure might be quite targeted, so the intended outcomes could be expressed in terms of seeking to address particular harms or behaviours on a service (e.g. reducing the risk of hate and terror content or child grooming on the service).

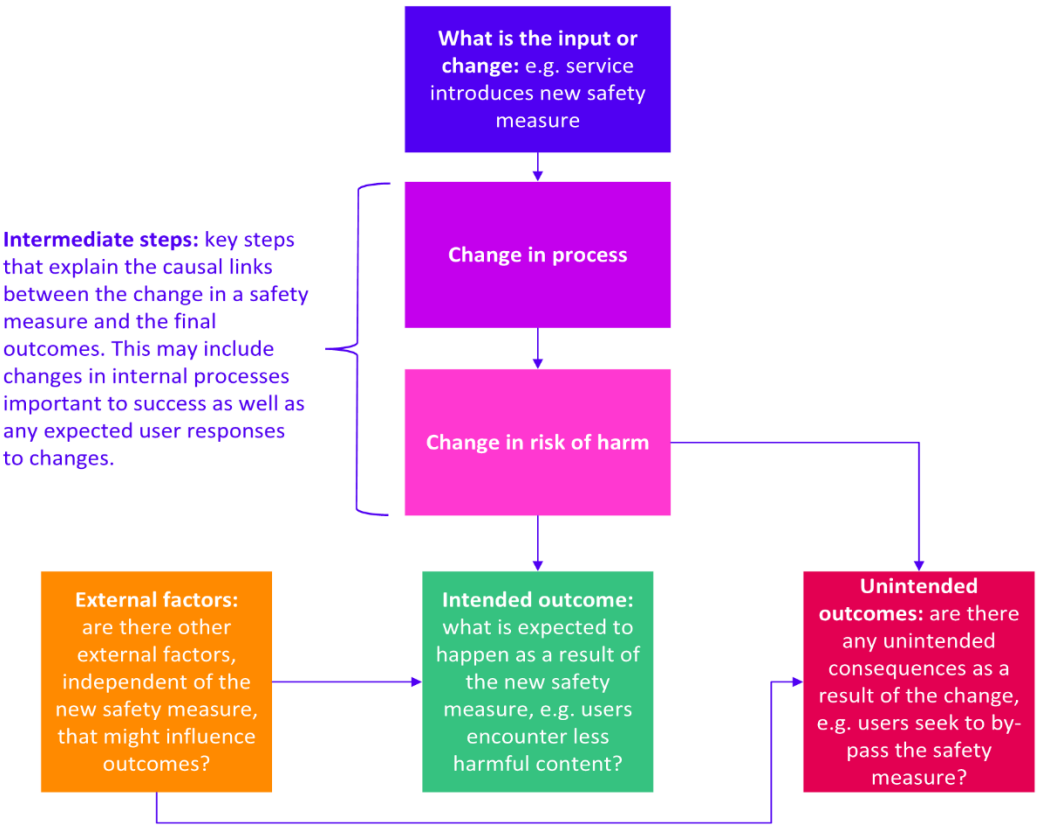
2.20 In other cases, the intended outcomes may relate to more cross-cutting safety measures (e.g. effective moderation systems to take down all known harmful content in a timely and appropriate manner).

Step 1b: Analysis of how the measure will have an impact

2.21 Once the objective of the safety measure is clear, the next step is to understand **how** it will achieve its intended effects. Evaluations often use an approach called a **Theory of Change** (TOC) to map out the causal links that describe how the introduction of policy interventions/activities lead to the final intended outcomes (see Figure 2 below). This is achieved by identifying the intermediate steps that need to be in place to reach the final outcomes. In carrying out this mapping, the TOC should identify underlying assumptions and uncertainties, risks and barriers to good outcomes. In addition, in the context of online

safety, safety measures often rely on a number of factors working together well and the service as a whole operating in a way consistent with safety objectives, and it is important to consider these interdependencies.

Figure 2: Example of a theory of change



Source: Ofcom

2.22 As shown in Figure 2 above, a TOC seeks to map out the intermediate steps between the introduction of a new safety measure and the final outcomes. In the context of online safety measures, it can be useful to consider this across different stages[5]:

a) Process stage: this initial stage starts by considering what safety measure has been (or is expected to be) implemented and how this occurs. This would consider not just the details of the safety measure, but the accompanying internal system and process changes that need to happen on the service for the safety measure to operate as intended. For example, a service introducing a new automated content classifier for hate speech may have to introduce a number of systems and processes, such as regularly updated databases that hold details of hate content.[6] But a TOC would also look at other

[5] The World Economic Forum has published related work in the Typology of Online Harms report. See page 5, World Economic Forum, August 2023, Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms, https://www3.weforum.org/docs/WEF_Typology_of_Online_Harms_2023.pdf [Accessed on 21 May 2024].

[6] For discussion of automated content classifiers, see Ofcom, March 2024, Evaluation in Online Safety:

supporting processes on the service that are needed for the new safety measure to work as intended (e.g. how the service's moderation system handles content flagged as potential hate speech to determine what content is available to users). In other cases, a process stage could involve a service changing how users can interact with the service or other safety measures made available to users (e.g. making user tools available to filter content). A process stage may therefore include consideration of how users are then meant to react to such changes (e.g. user take-up and use of a new safety tool).

b) Risk stage: this stage considers how the safety measure impacts the risk of harm to users. The risk of harm may be affected by the likelihood of users encountering harmful material; or related to the behaviours of users of the service (both those viewing content and those creating and uploading content) in response to a safety measure.[7] For example, improved detection should impact the prevalence of harmful content discoverable or served to end-users, which in turn should impact (reduce) the likelihood of users being exposed to harm.

c) Harms stage: this stage captures how a safety measure ultimately impacts the harm users might experience given changes to the risks of users encountering harmful material/behaviour on the service. For example, if a service is successful at reducing overall prevalence of harmful content, then this may feed through to fewer users encountering and engaging with that content and hence an overall reduction in harm.

2.23 The relationship between different stages needs to be considered carefully, including any interdependencies, feedback effects, external factors and unintended consequences. For example, while the safety measure may reduce risks of harm to a number users, an unintended impact could be greater harm for some users (e.g. due to users that seek to by-pass the safety measure and as a result end up encountering more harmful material). Therefore, the net effect on users will depend on the scale of both intended and unintended effects.

2.24 Even if a safety measure is expected to reduce harm experienced by users overall, it is still relevant to consider unintended impacts. For example, in response to safety measures such as enhanced transparency and sanctions against harmful material, content creators might alter the amount or severity of harmful content being uploaded to a service. But this could result in some content creators being restricted from expressing themselves or carrying out legitimate business in ways consistent with users' safety.

2.25 As well as overall levels of harm, it is important to consider the distribution of harm experienced by different types of user, especially those who are vulnerable. As such, the relationship between risks and harm outcomes may also need to consider how content can be experienced differently by different users. It may also be relevant to identify at this stage any external factors that that could impact on outcomes.

---

a discussion of hate speech classification and safety measures, Economic Discussion Paper Series Issue 9, https://www.ofcom.org.uk/__data/assets/pdf_file/0020/280217/Evaluation-in-Online-Safety,-a-discussion-of-hate-speech-classification-and-safety-measures.pdf

[7] For example, where the user views content, they could make use of user-empowerment safety tools to filter out harmful content. In addition, in response to service measures such as enhanced transparency and sanctions against harmful material, content creators might alter the amount or severity of harmful content being uploaded to a service.

2.26 In Part II we provide a detailed TOC as an illustrative example (related to evaluation of parental content controls) that elaborates on these points.

Step 2: Identification of success indicators and relevant metrics

2.27 Once the TOC is mapped out, it can be used to identify key success indicators and relevant metrics for the evaluation.

Step 2a: Identification of success indicators

2.28 Success indicators tell us whether the intended outcomes were, or are expected to be, achieved. These indicators are key outputs or effects that result from the measure. A success indicator can be a measure of success in terms of final intended outcomes, such as a reduction in the amount of harmful material viewed by children on a service. It can also relate to an intermediate step in the TOC, showing whether or not a safety measure is having the necessary impact on a particular part of the process. An example of such an intermediate success indicator is an increase in the speed with which harmful material is detected by moderators and removed from a service.

Step 2b: Identifying relevant metrics for each success indicator

2.29 In order to understand progress against success indicators, they need to be measurable. As such, suitable metrics should be identified for any given indicator.

2.30 In the evaluation of the effectiveness of online safety measures, we can think of metrics as falling under three main categories, relating to the three different stages of the TOC outlined above:

- **process metrics** focus on details around implementation of a safety measure (e.g. the functionalities and design of the safety measure) and the operational performance of a service's systems and processes. Process metrics do not attempt to directly measure the risk and experience of users encountering potentially harmful content nor the actual harm outcomes for users. Instead, they focus on the systems and processes that the service has in place and how these are working (e.g. KPIs showing that a service is reviewing reported content quickly and accurately).
- **risk metrics** focus on the risk of users encountering harmful material or behaviours on a service. This could include metrics of prevalence of harmful material on the service, but also the probability that users encounter and view that material, e.g. based on how and where it is made available on a service, or on what kinds of content creators are incentivised to create.
- **harm metrics** reflect overall change in harms experienced by users or wider society from exposure to harmful material or behaviours online. In this context, harms include those directly experienced by users (i.e. the user views content and experiences the harm themselves), but also more indirect societal harms (e.g. a user engaging with hate and terror content could lead to radicalisation of that individual with negative consequences for wider society). Harm metrics could also relate to the benefits that users miss out on if they spend less time online as a result of harmful material being there. For example, there are many positive well-

being benefits people can get from being online,8 but some users may choose to limit their online use in response to harms they have experienced (and/or due to their concerns about the risks of encountering harmful material).

2.31 A robust evaluation should also consider knock-on effects of a safety measure that could influence user outcomes, either positively or negatively. Therefore, it can be useful to identify **unintended outcomes metrics** separately. While, in some cases, unintended outcomes might be reflected in harm metrics, it is also useful to consider unintended outcomes arising from impacts in certain categories such as freedom of expression (i.e. takedown of non-harmful content), privacy, innovation and competition. Unintended consequences could also relate to a safety measure triggering some users to by-pass the measure on the service or seek out similar content on rival services.[9]

2.32 It is important that metrics, wherever possible, follow the S.M.A.R.T framework (specific, measurable, achievable, relevant and timebound). However, doing so is not without challenges. For example, measuring how harms manifest is inherently subjective, in the sense that the harm experienced by users may well vary depending on the vulnerability of those accessing content, the frequency and content of exposure, etc (we discuss this and other challenges at the end of Part I below).

2.33 The table below provides some examples of metrics related to the identification and removal of harmful content, broken down into the core categories of process, risk, harm and unintended outcomes.

[8] The benefits forgone include for example positive wellbeing benefits of online services due to knowledge acquisition, connecting with others, enjoyment and self-expression. See paragraph 3.20, Ofcom, 6 October 2021. Video-sharing platform guidance- Guidance for providers on measures to protect users from harmful material. https://www.ofcom.org.uk/__data/assets/pdf_file/0015/226302/vsp-harms-guidance.pdf [Accessed 31 January 2024].

[9] In some cases, unintended effects could be detected through harm, risk or process metrics. For example, an unintended impact of a safety measure could be that it triggers some users to by-pass the measure. This might be indicated by metrics that show users encountered an increased number of pieces of harmful content. However, it may be difficult to identify this impact using overall metrics, as the net impact of a safety measure could be that it reduces the amount of harmful content users encounter.

Table 1: Example metrics by category

Metric category	Metric sub-category	Metric
Process metrics[10]	Methods of identification	Details of technologies or methods to identify harmful content that services have implemented
	Volume of harmful content identified and removed	Volume of suspected harmful content identified by different safety measures or technologies Volume of content created by repeat offenders that is removed
Risk metrics	Prevalence of harmful content	Volume of harmful content found when randomly sampling content on the service Share of harmful content out of the top 10k most viewed pieces of content
	Exposure to harmful content	Number of views Time elapsed before harmful content is taken down Number of user reports of harmful content
Harm metrics	User experiences of online services	Subjective wellbeing of people making use of relevant services, including vulnerable users
	Specific harms reported	Number of people reporting specific harms experienced on online services e.g. via survey responses or reports to a service
		Number of reported/upheld reports of hate speech content against people based on their protected characteristics
	Benefits forgone	People reporting that they have reduced the number of online services they visit or are spending less time using them due to harms experienced
Unintended consequences metrics	Freedom of expression	Percentage of content identified as violative and removed that is successfully appealed by the end user (over-enforcement of content)

[10] In developing this discussion paper, we examined the different types of metrics either published and used internally by different online services. We found that the largest number related to process metrics.

Stage 3: Assessing the outcomes

Track, monitor, assess outcome metrics

2.34 Once relevant metrics are identified and collected, the next step is to interpret those metrics to understand what they tell us about the effectiveness of the measure that is being evaluated. There are a range of ways in which this can be done, with varying types and depth of insight possible depending on the specific context of the evaluation. We summarise below four different methodologies that can be employed - quantitative comparisons and pre/post analysis; econometric techniques; A/B testing; and behavioural science approaches.[11] For each approach, we briefly explain what is involved and provide some considerations as to when they might be most appropriately used.

Quantitative comparisons and analysis

2.35 One of the main ways to assess outcomes is simply to track changes in identified metrics over time. This type of analysis can be based on data collected directly by services, via user surveys and from other sources. It can look at broad trends or detect discontinuities in the data (e.g. a sudden jump either up or down in reported harmful content on a service).

2.36 A high-level analysis like this has some limitations, such as difficulties in making comparison across services and accounting for external factors. In addition, drawing causal links between safety measures and impacts on users' experience of harm can be complex. For example, a service could introduce a well-designed safety measure that improves detection and reporting of harmful material, but an unrelated external event (e.g. an election or war) could still lead to an overall increase in the number of uploads of harmful content on the service. Not controlling for such events could result in wrong conclusions about the efficacy of the measure.

2.37 Despite some of these drawbacks, relatively straightforward quantitative comparisons can be helpful in establishing a baseline and assessing whether outcomes are moving as expected. It is often the first step in an evaluation and can inform more sophisticated techniques by narrowing down the set of targeted questions and data required.

Econometric techniques

2.38 Econometric techniques can be used to better understand the relationship between different variables (such as the introduction of a new safety measure and levels of harm experienced by users). Going beyond the simple comparisons discussed above, econometric analysis can help to isolate the impact of a specific safety measure on an outcome (for example, volume of harmful content). It does this by controlling, as appropriate, for other factors (confounding factors) that could also influence that outcome.

2.39 Because it is often trying to infer causality (the degree to which one change is affecting a particular outcome), econometric analysis can be complex. It often requires rich data across services, users, or content and often over time. Because of the challenges of conducting

[11] A detailed description of the range of evaluation tools available and used can be found in Annex A, HM Treasury, 2022. The Green book- Central Government Guidance on Appraisal and Evaluation https://assets.publishing.service.gov.uk/media/623d99f5e90e075f14254676/Green_Book_2022.pd [Accessed on 31 January 2024].

robust econometric analysis, the choice of a causal study to assess the impact of a safety measure will depend on factors such as:

- the importance, novelty, scale and learning potential of the safety measure; and
- technical feasibility, for example availability and reliability of suitable data, and time since introduction of the safety measure.

Box 1: Online false information case study

Ofcom has used econometrics across a wide range of regulatory areas,[12] including for online safety, such as in Ofcom's report on understanding online false information in the UK.[13] The project involved web scraping both trustworthy information sites and false information sites categorised by NewsGuard,[14] and then running correlational studies between the two types of sites with a number of variables. The relationships that were explored included the correlation between a trusted and a non-trusted site with the number of visits, the age group of users, the gender of users, and the topic of the sites. Using this data and controlling for other trends, we have developed an approximation for the characteristics of the main visitors of false information sites.

2.40 As econometric analysis will usually require large datasets - often panel or timeseries data - planning in advance for an ex-post evaluation will tend to make it easier and more cost effective. This is because it will enable the identification and collection of the required data in a timely manner.

A/B testing

2.41 A/B testing, also known as split testing, can be used to test the effectiveness of a safety measure or variations of it. The idea is to split users into groups and apply a safety measure to one group of randomly selected users (A) and compare the results to a control group (B) that does not have access to the safety measure.

2.42 A/B testing is commonly used by online services for example to test new products or designs of their site to improve commercial outcomes.[15] In the context of testing safety measures, there can be drawbacks to A/B testing as it could potentially entail withholding a safety measure from one group of users, which could lead to some people in that group experiencing harm. If this poses a risk, A/B testing may be better suited to test different design features of a safety measure to see what works best. This concern can also be overcome where data is available prior to a safety measure being introduced. In this case, A/B testing can be used to compare outcomes for all users when the measure was not in

[12] For example, we conducted an econometric study on the evidence of the impact of mobile consolidation, including on investment. Ofcom, January 2021. Discussion Paper: Market Structure, investment and quality in the mobile industry. https://www.ofcom.org.uk/research-and-data/economics-discussion-papers/mobile-market-consolidation [Accessed 31 January 2024].

[13] Ofcom. 27 January 2021. Understanding online false information in the UK https://www.ofcom.org.uk/research-and-data/economics-discussion-papers/understanding-online-false-information-in-the-uk [Accessed 06 March 2024]

[14] The analysis used third party data from NewsGuard and therefore their standard of classification of false information.

[15] See for example, Big Commerce Blog, A Step-By-Step Guide to Effective Ecommerce A/B Testing: https://www.bigcommerce.com/articles/ecommerce/ab-testing/ [Site accessed 17 May 2024].

place (the before group) with outcomes for all users after the measure was introduced (the after group).

Behavioural techniques

2.43 Behavioural economics is a branch of economic analysis that can help predict and understand how users might behave in practice when confronted with particular scenarios. In this context, behavioural techniques can provide an insight into how effective safety measures are likely to be in reducing user harm, or into understanding why users have responded to safety measures in a particular way.

2.44 People's decision-making will not always be rational (in a strict economic sense), as it will be influenced by factors like cognitive limitations and biases. Behavioural analysis uses tools like randomised control trials (RCTs) to understand how users might respond to a safety measure to help us evaluate not just "what" the impact is likely to be but also "how" it is likely to come about and "why"' users have behaved in a certain way.

2.45 Behavioural audits are another tool that could be harnessed to support evaluating the effectiveness of safety measures. A behavioural audit involves the systematic documentation of how a user is likely to behave across a given set of features or functionalities. They can be used to identify features of the online choice architecture that warrant further investigation in terms of their impact on user behaviour or to compare the "before" and "after" of changes to service functionality.

2.46 Additionally, a behavioural model which can be used to assess different aspects of user behaviour is the COM-B model.[16] In the context of online safety, this model can be used to assess the Capability, Opportunity and Motivation of users to engage effectively with safety measures and, for instance, helps to identify barriers to understand why an online safety measure did not work. In Part II, an example of the COM-B model is included to illustrate this.

Box 2: Behavioural insights for online safety

Ofcom has a programme of research using online RCTs to explore how users might respond to a safety measure, see for example: Behavioural insights for online safety.

This research programme has utilised realistic mock-ups of VSPs and social media services to test the effectiveness of different behavioural techniques (e.g. 'nudges' and 'boosts') based on how users engage with safety measures. This research has generated important insights into users' online behaviour through testing users' behaviour in a research environment. That is, research participants were aware they were taking part in a trial (even if they didn't know what was being tested) and that they were being observed.

Using a mock-up also means that we need to be cautious about how far we can extrapolate from our research to real-life situations. We consider that the results of our behavioural experimental research can give an indication of the 'direction of travel' from the measures we tested rather than precise quantification of the magnitude of any impact.

[16] Michie, S., van Stralen, M.M., and West R. (2011) The behaviour change wheel: a new method for characterising and designing behaviour change interventions. See page 21, Ofcom, July 2023, Discussion Paper: Behavioural insights for online safety: understanding the impact of video sharing platform (VSP) design on user behaviour.

Stage 4: Evaluate and iterate

2.47 Once the analysis has been conducted and conclusions have been drawn about the effectiveness of the safety measure, these conclusions can be used to refine, improve or replace safety measures. For example, where existing safety measures are shown to be ineffective, services could adapt, remove or amend them and/or implement further measures based on the risks posed by their services. Safety measures that are found to work particularly well may serve as a best practice example for the sector and drive innovation of new techniques to effectively protect users.

2.48 One lesson from the field of behavioural insights is that often quite subtle changes in the design of user tools can alter user behaviours, sometimes in predictable ways, but there is still scope for unexpected user responses. The approach to evaluation in this paper can support a degree of experimentation by services to both learn and share lessons on what designs might work better in different circumstances.

Challenges and wider considerations around metrics

2.49 We have suggested above some key steps for an effective evaluation. There may be a number of challenges and wider considerations when conducting such evaluations and gathering data. This is particularly likely when evaluating final outcomes that deal with sometimes quite subjective concepts such as harms and which can be influenced by a range of other factors. In these circumstances, a number of metrics are likely to be required to make sense of outcomes and to determine what is driving them.

2.50 A range of data sources will often be needed to conduct meaningful evaluations. Services will need to collect and retain data from intermediate internal processes, processes they may have outsourced (e.g. moderation) and final outcomes. Services may also want to consider opportunities to conduct user surveys to understand their users' experiences or use data from third-party providers. We discuss the challenges and wider considerations around metrics below, and a further discussion of different data sources, including considerations and certain limitations, is available in Annex 5.

2.51 **Need for proportionality and prioritisation:** Any evaluation will need to be proportionate to the risks, benefits, and costs of the relevant online safety measures.[17] While a TOC should be comprehensive, setting out all the intermediate steps and potential outputs and outcomes, it does not mean every step in the TOC has to (or can) be measured. The choice of metrics and tools used to analyse outcomes may need to consider time, resources, and priorities (Annex 3 provides more details on possible relevant factors to prioritise metrics).

2.52 **Risk of Goodhart's Law:** The idea behind Goodhart's Law is that once you target certain metrics for measurement (for example to inform policy) that metric may no longer be a useful measure.[18] In the context of online safety, one risk is that regulated services may seek to focus their efforts on "performing" against certain metrics visible to regulators, rather than using them to genuinely inform and improve safety. For example, if we were concerned about the amount of time moderators were taking to remove content, because we thought it was reflective of insufficient training and resources, we could measure "average takedown times". However, this could prompt services to impose stricter time allowances for

[17] As set out in the Treasury's Green and Magenta Books, planning and provision of resources for monitoring and evaluation should be proportionate when judged against the costs, benefits and risks of a proposal both to society and the public sector.

[18] The four flavours of Goodhart's Law.

moderators to review each piece of content rather than investing in more moderator training and resources. In turn, if moderators spend less time reviewing content, this could lead to worse safety outcomes if harmful content were missed, or moderators felt rushed to make a decision. It is possible to overcome this issue, for example by looking at metrics in combination, such as assessing both the timeliness *and* accuracy of moderation.

2.53 **Metrics may only capture known and measurable outcomes:** The ways in which services identify and capture details of the amount of harmful content may vary, which may create challenges in creating reliable metrics. For example, some services generate prevalence metrics that attempt to show the proportion of harmful content relative to all content available on their service. Low rates of prevalence could indicate that a service is finding and removing content and has effective safety measures in place. However, the prevalence metric might rely only on content the service knows about due to reports generated by its users or content it detects. Rates of user reporting could be very low if a service's reporting tools are hard to find and use or if the service's definition of harmful material is too narrow. In this case, a low prevalence may partially reflect inadequacies in the service's reporting or detection rather than successful removal of harmful content.[19] Therefore, in this example, a range of intermediate metrics may be necessary to show reporting and detection are working well.[20]

2.54 **Use of proxy metrics**: There may be challenges in measuring certain harm outcomes directly, which point to the use of proxy metrics. For example, if there were issues gathering data to measure harms experienced by users, a possible proxy metric might refer to exposure of users to content labelled as harmful and/or the number of users that viewed said content before it was removed by a service. However, there are limitations to the use of proxy metrics and care should be taken with their interpretation. For example, exposure metrics say little about the actual personal harm (e.g. physical or mental) or social harms experienced. Not every user who sees potentially harmful content will experience actual harm - it can depend on a user's individual level of vulnerability, and some will experience it only after repeated exposure.

2.55 **Consistency of metrics and of measurement across services**: One issue with evaluation that uses comparison of metrics is that they are not always compiled on a standardised basis (for example, how services measure their monthly active users).[21] This is an important point to note from the perspective of regulatory evaluation where Ofcom may seek to compare effectiveness of measures across services. Standardisation is important where the goal is to aid comparisons across services, as it can provide an important reference point to certain metrics. But standardisation might be less important where an evaluation is service

[19] Similar considerations apply if a service own systems and processes to detect and review harmful content do not work well.

[20] Some services also seek to measure the amount of harmful content by looking at a relevant sample of content available to users on their site. For example, the Integrity Institute regularly tracks information based on transparency reports of some of the largest online services such as Facebook, Instagram and YouTube. It notes that these services publish prevalence metrics based on sampling methods, but based content with the most impressions ("Top-N" lists). The Integrity Institute notes however that these samples reflect content that is most successful and recommended by the services, so will not necessarily represent an unbiased sample, see: https://integrityinstitute.org/widely-viewed-content-analysis-tracking-dashboard [Accessed 26 April 2024]

[21] For example, a service could provide total levels of user-reported harmful content, but one service might count multiple reports by different users of the same content whereas another service might only count unique pieces of content reported. In addition, in categorising potentially harmful content, one service might rely on the reason given by the user and another on the moderator decision.

specific.[22] Different measurement approaches might not necessarily be 'wrong' and in some cases one approach may be better suited to answer different evaluations. Some commentators, such as Daphne Keller, have also highlighted that attempting to standardise across services also risks nudging services to standardise their actual policies and practices (e.g. defining consistently what counts as cyber-bullying), saving expense but reducing pluralism and technical diversity of online services.[23] Therefore, we need to be aware of certain trade-offs in the design of evaluation metrics.

[22] In this context, consistency of measurement will still be important when looking at outcomes over time.
[23] Statement of Daphne Keller, 5 May 2022. Before the United States Senate Committee on the Judiciary, Subcommittee on Subcommittee on Privacy, Technology and the Law, Hearing on Platform Transparency: Understanding the Impact of Social Media. https://www.judiciary.senate.gov/imo/media/doc/Keller%20Testimony1.pdf [Accessed 31 January 2024].

3. PART II – Evaluating the effectiveness of Parental Content Controls

Introduction

3.1 In Part II of this paper, we provide a worked example applying the evaluation steps, based on the TOC approach discussed in Part I. We have used an example of a child safety mitigation (parental content controls) on VSPs to demonstrate how the evaluation framework discussed in Part I could be used to test its effectiveness in protecting children from harmful content.

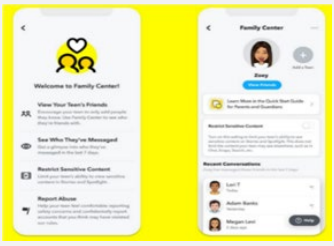
3.2 This worked example is included for illustrative purposes only. It is not intended to provide an assessment of any particular platform's safety measures, specific recommendations of metrics a platform needs to collect, nor does it represent the only way in which Ofcom or services should seek to conduct an evaluation. Our intention is to illustrate how applying a TOC can help identify key implementation steps and to give a flavour of the considerations services and Ofcom may need to take into account.

3.3 Parental content controls aim to protect children by empowering parents to make more decisions about the type of material they do not want their child to see or how their child might be able to interact with a service (see Box 3 for more detail).

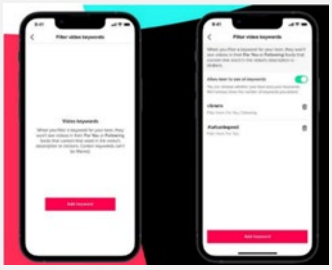
3.4 As with many safety measures, parental content controls used in isolation of other measures cannot provide a panacea solution to protecting children online. Comprehensive online child protection is likely to require effective use of other safety measures, as set out in our VSP Child Safety Report 2023.

Box 3: What are parental content controls?

Parental content controls are a form of user empowerment tool, often one part of a wider range of tools or strategies that parents may use to help keep their children safe online. From our regulation of VSPs, we know that the availability and types of parental content control tools differ between the services. Some services who don't target younger demographics often do not have these tools at all. In our Child Safety Report 2023, we noted that provide some sort of controls (e.g. Snap and TikTok since 2022 and 2020 respectively) typically have 'centres' which host a variety of available controls or tools and information for families.



Screenshot of parent's view of Snapchat's Family Centre. See: Figure 2.5, Ofcom, How video-sharing platforms (VSPs) protect children from encountering harmful videos (ofcom.org.uk)



Screenshot of parent's view of TikTok's Parent Control centre. See: Figure 2.4, Ofcom, How video-sharing platforms (VSPs) protect children from encountering harmful videos (ofcom.org.uk)

The types of parental content controls typically used can generally be broken down into:

1. Active mediation tools - which encourage or provide a service for discussion between the parent and child, including:

- Support pages, learning resources and guidance for parents and potential mediation prompts to encourage parents to discuss with their child their online life.

2. Restrictive mediation tools - which involve a parent regulating their child's online experience and activity, including:

- Daily screen time limits - that set a limit on the time that a child can spend on the app.
- Restricted mode - to restrict a child's exposure to specific pieces of content.
- Discoverability tools - to give parents the ability to set the child's account to private.
- Liked video and comment control - to restrict what type of user accounts can like or comment on the child's account.
- Keyword search filters - to enable parents to restrict which terms a child can search for.
- Direct message controls - to restrict who can send messages to a child (i.e. known contacts).

3. Monitoring tools - involve the parent monitoring their child's online activity, including:

- Direct message monitoring - to allow parents to monitor other users their child is contacting (often not the content of the message, rather just the name of the contact).
- Friends and liked subject list monitoring - to allow parents to monitor who their child is "friends" with and who is liking their comments.

3.5 In our example below, we have focused on the following parental content control tools:

- A service designs a parental content control tool which gives the option for a responsible adult to pair to their child's account.
- This pairing then gives the parent options that give some control of the content their child can see, either through the parent setting 'filter video' keywords or applying a 'safe' mode.

Step 1a: Set out the purpose of implementing parental content controls

3.6 The first step is to identify the purpose of implementing parental content controls. In this case, the parental content control tool is designed to decrease the likelihood that a child encounters content that might be harmful for them or content that a parent deems inappropriate for that child while they use the service. Hence, a measurable outcome would be that children on a service and that are known to have parental tools active encounter fewer pieces of harmful content.

3.7 The level of control and type of content that the parent chooses to filter out is likely to vary depending on a child's age, as well as additional factors unique to that individual.[24] In designing the tool, services should balance the dual aims of protecting children from harm, while also upholding their fundamental rights to privacy and freedom of expression. Generally, children's autonomy develops over time, as does what is in their best interests, and so the approach taken by parents will likely need to adapt with the age of the child. For example, as the child gets older the parent might consider transitioning from restrictive controls to mediation-based tools. Excessive restrictions, even where well-intentioned, could remove a child's ability to benefit from online participation. This could include benefits from the use of social platforms through building knowledge, connection, enjoyment, and expression.[25] Any evaluation should consider the entire set of intended and unintended outcomes and a TOC will help in setting this out.

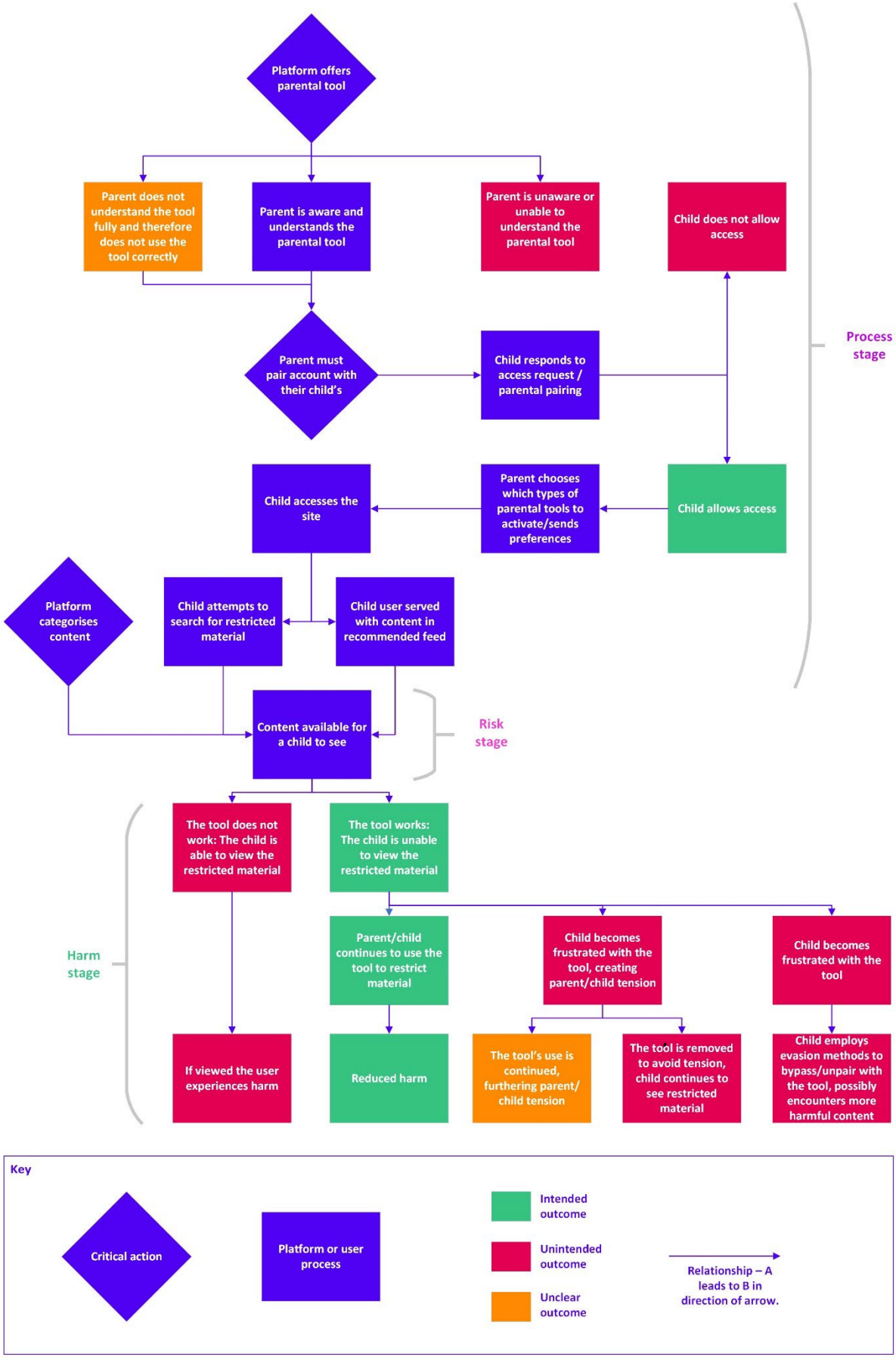
Step 1b: Analyse how parental content controls will reduce child users' exposure to harmful material

3.8 The TOC presented in Figure 3 below sets out how parental content controls might affect what child users see and considers both intended and unintended outcomes.

[24] 13-17 year older users should already be protected from restricted material. Therefore, parental tools focus on material that goes beyond the scope of restricted material. We have called this material perceived harmful material. Importantly perceived material may be harmful for one user but not another.

[25] See page 22, ICC, March 2021. Research on Protection of Minors: A literature review and interconnected frameworks. Implications for VSP regulation and beyond. https://www.ofcom.org.uk/__data/assets/pdf_file/0023/216491/uel-report-protection-of-minors.pdf [Accessed on 31 January 2024].

Figure 3 – Theory of change associated with parental content control tools



Source: Ofcom

3.9 The above TOC can broadly be described in terms of the following key stages: [26]

Process stage:

- **Availability of the tool**: what are the control features and options that have been made available to parents? [27]
- **Access and take-up**: are parents aware of the tool and is it easy to use, is the parent required to have an account on the service themselves, and do parents understand the nature of the service and the content and features that may appear within it? Without this knowledge and ease of access there is a risk that parents might quit or disengage with the process. Take-up can also be affected by the child's behaviour, which is often dependent on the parent-child relationship.
- **Service's processes:** for the tool to be effective, the service's supporting systems and processes must also be well designed (i.e. accurate content labelling, classification and detection of mislabelled content).

Risk stage:

- **Content available to the child:** if tools are available in an accessible manner, parents make use of them effectively and service processes work well, then there should be a lower risk of the child seeing harmful material.

Harm stage:

- **Intended outcomes:** if the controls are put in place and they lead to restrictions in harmful content, this should lead to the child encountering fewer instances of harmful material and as a result experiencing less harm.

Unintended outcomes:

- Unintended outcomes might include parents engaging with the tools incorrectly, curtailing the child's right to freedom of expression, or the child seeking to evade the tool by switching to a less regulated service.

3.10 As noted, our worked example only refers to a limited set of parental content controls. To fully understand the outcomes, an expanded TOC might need to consider other types of parental tools and how these fit within the overall suite of other potential safety measures and any other key interdependencies. [28]

---

[26] One service process could fit into a number of different stages. For example, the "parental user requires parental pairing with a child" would fit into both the initial take up and engagement, which could be affected by technical ease of use and the starting parent-child relationship. But, also once the parental content controls occur, this could affect the parent child relationship and subsequently lead to a negative response by the child such as unpairing of their account.

[27] Although we might still be interested to understand outcomes for users on services without parental tools relative to those that make them available.

[28] For example, most VSPs currently rely on self-certification of age by users, which can lead to underage children signing up to platforms and/or children accessing more mature content. Availability of parental control tools rely on the account being recognised as child owned. If a child has misrepresented their age as 18+ through self-certification, the service will not offer the option to connect to a parent and will not benefit from the protections. Parental content controls are therefore dependent on understanding the correct age of the child.

Step 2a: Identify success indicators

3.11 Based on the key stages identified in the TOC example above, we have identified six key success indicators that could help assess the effectiveness of parental content controls.

Availability

3.12 Understanding which tools services have to offer is pivotal to the evaluation process and is likely to be ongoing as services introduce, update, and withdraw different features of their tools.

Access and take-up

3.13 After a service has made a parental content control tool available, people need to use it for it to have any effect on user safety. Even if a tool is effective when used, without reasonable levels of uptake, it will have a limited impact.

3.14 The take-up of parental content controls is likely to be influenced by a range of factors, including users' motivation, awareness and the ease of use of locating and setting up the tools.

3.15 Ofcom research has shown that 58% of parents perceived VSPs to have some form of parental content controls in place.[29] There is limited data on uptake and usage of parental tools on VSPs where they have been made available. However, as reported in our VSP child protection report[30] Snap told us that use of its Family Centre represented 1-2% of its total under-18 daily active users. We noted in the report that, while users may be more likely to engage with parental controls if they perceive them to be effective, uptake rates do not necessarily indicate the effectiveness of a parental controls system itself. The lack of use could be due to an absence of organic demand from parents and/or children, or due to other factors.

3.16 Where take-up remains low, there could be merit in further examining the causes. Given the user bases of larger services is high in absolute terms, further increases in uptake could still impact a large number of accounts, even if not all parents or children chose to use these tools.

3.17 In addition to initial set-up by users, it is important to assess the extent to which tools remain active. As set out in our TOC, the activation of the parental content control tool might rely on continued pairing of the parent and child's accounts. The child may initially refuse pairing or unpair their account later on, for example as a result of updates to parental content control tools or in feeling an increased sense of autonomy as they get older.

3.18 As a longer-term goal, although perhaps lower priority, we might also want to understand why take-up rates might vary accounting for different factors (for example, increases in advertising affecting awareness of their availability, or wider societal changes driving changes in attitudes to online safety issues).

---

[29] See page 29, Ofcom produced by YouGov, 28 October 2023. VSP Tracker Wave 4. https://www.ofcom.org.uk/__data/assets/pdf_file/0027/272259/vsp-tracker-wave-3-4-chart-pack.pdf [Accessed on 31 January 2024].

[30] See page 16, Ofcom, 14 December 2023. How video-sharing platforms (VSPs) protect children from encountering harmful videos. https://www.ofcom.org.uk/__data/assets/pdf_file/0020/273224/vsp-child-safety-report.pdf [Accessed on 31 January 2024].

Ease of use and user behaviour

3.19 For parental content controls to be effective, users will need to find them easy to use. The design of the tools will affect whether parents engage with the tools appropriately. An indicator of this can be the extent to which there remains ongoing engagement and use of the tools once parents and children have signed up to them.

3.20 Important factors here include the technical ease of use of the parental content controls and whether realistic expectations are placed on parents. Tools may give parents lots of choices to customise and control their child's experiences, but this could risk over-burdening them with the responsibility to make the right choices and be aware of a significant range of harms.

3.21 When assessing the technical ease of use of parental content control tools and the potential challenges users could face, it may be relevant to see whether the tool adheres to principles of good design. This could be done via user views or tracking user interactions with the tools, user surveys or the COM-B model. Below is an example using the COM-B model:

- **Capability:** Do parents and children have the technical capability, levels of media literacy and sufficient understanding of under-18 online culture to set up parental tools to make them work effectively?[31]
- **Opportunity:** Do parents have enough time to invest in learning about how to use the tools effectively and how the service works and to regularly engage with the tools (if the design of the tool requires this)?
- **Motivation:** If a tool requires too much effort or frustrates either a parent or child, it may result in the parent giving up. Specific points of failure include setting up the tool, particularly if this is an intensive process and if the tool needs to be revisited to remain effective.

3.22 One challenge in measuring ease of use is that there may be significant differences between users' needs based on their particular COM-B. In this context, it would be difficult to ascertain what a reasonable number of customisable settings might be for a tool. Additionally, a tool that offers multiple default settings may end up restricting the parent's choice and result in a poor fit between the settings in place and the affected child user.

Key service processes

3.23 As noted in our TOC, once a parent-child account has been paired and the parent has set their preferences (possibly in discussion with their child), these need to be reflected in the content their child can see. If the service does not have effective processes to categorise and classify content, then this risks content that the parent would not want their child to see still being available to the child.

3.24 Due to other safety measures services may have in place, services may already seek to classify and restrict content. Therefore, some of these parental preferences for content restrictions may overlap with content the service attempts to target under existing protection measures. In other cases, some parental preferences may relate to content that

[31] For example, if the tool offers keyword blocking options then a parent would potentially have to locate the parental control function and then select keywords for their child's account, so that when a child tries to search this key term they receive a blank result. Even with some technical knowledge, many parents may be unable to comprehensively list keywords to block for their child, as this process would require an understanding of coded language and current teenager trends. However, a parent could potentially be aided or guided if the tool provided some keyword suggestions.

is more borderline in terms of whether it falls within the scope of illegal or harmful content set out in the respective legislation. Parents may do this to prevent a child from experiencing a certain type of harm to which they are particularly vulnerable.

Intended outcomes

3.25 The main intended outcome of parental content controls would be a reduction in harmful content encountered by the child. Another benefit of these tools is that they also encourage wider dialogue between the parent and child around harmful material.

3.26 We would ideally assess the effectiveness of parental content control tools based on the incremental benefits of parental content control tools as an additional layer of safety. This would need to be carefully measured given that these tools sit alongside other existing measures to reduce the risk of harm (applied to all known child accounts or user accounts more generally). In this respect, we would ideally want to control for content that is unavailable to child users due to other safety measures.[32] However, determining whether harmful content was or was not available due to parental content controls or some other safety measure may be challenging.

Unintended outcomes

3.27 There are a number of unintended consequences that could arise from the use of parental content control tools. In some cases, these unintended consequences might reduce the benefits the child gets from using the service, and in some scenarios could even lead to more harm.

3.28 Family structures and parental engagement is varied. This may result in some parents using the tools in an overly restrictive way.

3.29 Even where these tools are used in the context of a trusting relationship, the parental content controls may, as a by-product, inadvertently restrict access to content or services with positive benefits, such as reduced access to educational content and other factors such as social engagement and wellbeing.

3.30 Other unintended outcomes could arise due to a child's response to their parent's use of parental content control tools. Children may dislike parental involvement in their online life. They may feel their privacy or right to participate has been infringed, leading to attempts to bypass the tools via new accounts, unpairing or switching to alternative services. This could have the unintended consequence of exposing the child to more harmful content. This might arise for example if the child moves to alternative services with fewer safety measures in place; and/or because the parent may have reduced awareness and oversight of the child's use. Issues such as those around the autonomy, privacy, and freedom of the child are likely to become more prominent as the child ages.

3.31 Parents may also have unrealistic expectations of parental tools, believing that they are a catch-all to remove harmful material. Therefore, parents may not engage in the other

[32] Some safety measures aimed at protecting children do not always entail content being made unavailable. For example, in our Child Safety Report, we noted that Twitch applies maturity warning labels, but still allows under 18s to view potentially mature content. Therefore, while both a maturity label and parental tool might apply to similar content, they would differ in terms of the levels of restrictions applied, which in turn could impact levels of harm experienced by the child. See page 11, Ofcom, 14 December 2023. How video-sharing platforms (VSPs) protect children from encountering harmful videos. https://www.ofcom.org.uk/__data/assets/pdf_file/0020/273224/vsp-child-safety-report.pdf [Accessed on 21 May 2024].

processes needed to ensure that their child is safe online, such as dialogue and online education. Some research has suggested that improving a child's digital resilience and awareness may be as important to harm reduction, and potentially more so, than the introduction of filters or blocking.[33]

Step 2b: Identify relevant metrics

3.32 Based on the success indicators, and the discussion above on some key questions, we identify some illustrative candidate metrics in Table 2 that, where available, could be used in evaluation of parental content controls. To develop a comprehensive picture, a range of metrics will be required to make sense of the impacts of the parental content control tools, and to understand how the tools could be made more effective over time. The metrics listed in Table 2 are not intended to be a required, complete or final view of metrics services (or Ofcom) should use.[34]

3.33 Both the success indicators and metrics can be prioritised according to the risks, size and needs of the specific evaluation (see Annex 3 for further details on prioritisation).[35]

Table 2: Illustrative candidate metrics for each success indicator

Success Indicator	Key questions	Illustrative candidate metrics
Availability	What is the nature of parental content control tools, if any, offered by a service?	1. Main options available by parental tool by type (e.g. restrictions on search, recommendation, time restrictions etc.). 2. When parental tools were first made available to users. 3. Parts of the service to which parental content control tools can be applied.

[33] See page 63, ICC, March 2021. Research on Protection of Minors: A literature review and interconnected frameworks. Implications for VSP regulation and beyond. https://www.ofcom.org.uk/__data/assets/pdf_file/0023/216491/uel-report-protection-of-minors.pdf [Accessed on 31/0/1/2024].

[34] In addition, for each candidate metric would need to be specified more fully to include relevant units, timeframe for collection etc.

[35] In Annex 3, we highlight the main factors that could be used to prioritise metrics. These are salience for measuring effectiveness of the safety measures, ease of analysis and robustness, and, practical, cost effective, and appropriate.

Success Indicator	Key questions	Illustrative candidate metrics
Take up and engagement	What are levels of take-up and ongoing engagement with parental content control tools?	1. Total number of child accounts by child's known age. 2. Total number of child accounts paired to a parent account. 3. Number of new child accounts paired per month by child's known age. 4. Total number of unpaired accounts per month, including by: child's known age; whether child or parent initiated. 5. Tenure of parental tools (i.e. how long in days/weeks the tool remained paired). 6. Number of paired accounts where the parent has active settings. 7. Number of parents that initiated set-up of parental content control tools but did not complete the process.
Ease of use and user behaviours	Based on the design of the tool, is there ongoing engagement and use of the tools reflecting ease of use and expected user behaviours?	1. Perceptions of users on whether parental content controls on services are easy to use. 2. Average time parents spend using the tools. 3. Percentage of parent accounts that have changed the configuration of the parental content controls. 4. Average number of parental content control settings changed by a parental account/for a child account on the service. 5. Average number of active parental content control settings for a parental/child account on the service. 6. Frequency with which parents review or revisit tools.
Service's process	Do services successfully categorise content to reflect parental settings?	1. Method(s) through which content is categorised (e.g., through automated tools, creator labelling or user reporting). 2. Number of unique keywords input into parental content control tool and proportions that match / do not match service classifiers. 3. Number of unique videos/content not made available by different surfaces on the service due to parental settings versus other safety measures.

Success Indicator	Key questions	Illustrative candidate metrics
Intended outcomes	How successful are the tools at preventing child users from viewing perceived harmful material?	1.Number of child accounts with parental supervision tools who viewed content labelled as 'Restricted Material'/ 'Perceived harmful' content. 2. Views/impressions from child accounts to content labelled as 'Restricted Material'/ 'Perceived harmful' content. 3. Percentage of children who reported (via surveys) having seen restricted material on the service, by those with/without parental content control tools active. 4. Reports to the service of violative content from child accounts by those with/without parental content control tools active.
Unintended outcomes	Does the pairing impact the way children used their account in a way that leads to more harmful experiences?	1. Survey of parents who reported that the tool led to increased parent-child tension 2. Count of instances when a child account unpaired/unlinked from a parental account. 3. Number of children who created new accounts without a parental content control tool active. 4. Survey of children who report switching to services with no parental content control tools. 5. Monthly traffic to out-of-scope services offering similar services with no parental content controls.

Source: Ofcom

Step 3: Assess the outcomes

3.34 Once relevant metrics have been identified for measuring success indicators, data will need to be collected against the metrics.

3.35 To give the best picture about the impact of implementing parental content control tools certain data collection may need to take place over a period of time or at certain milestone points. For example, data can be collected before the introduction of the parental content control tools, at periodic intervals following their implementation, and following any iterations e.g. after updating any bugs.

3.36 When the relevant data has been collected, the next stage is to determine which methodology to use to best assess the metrics and draw conclusions about the effectiveness of the parental content control tools. In some cases, it might be straightforward to track individual metrics to draw conclusions. For example, the metric 'total number of child accounts paired to a parent account' (from Table 2) could be collected over a period of time to understand the take up of parental content control tools over time. However, typically, a range of metrics and, if necessary, analytical methods and tools, may be needed to provide more context and address challenges and limitations. We illustrate below considerations that might need to be taken into account when applying different evaluation techniques in the case of parental content control tools.

Quantitative comparisons based on service's data

3.37 There are likely to be a number of challenges and considerations when using quantitative comparisons alone for metrics aimed at capturing the contribution parental content control tools might make to reducing children's exposure to harmful content. Difficulties could include:

- **User reporting difficulties:** One method of measuring the prevalence of perceived harmful material is relying on detection and/or user reporting of that content via service reporting tools. Difficulties could arise from lack of awareness about the reporting tools themselves (users may not know where to find them). In addition, content that is causing harm but is not recognised as such (as perception of harmful content may vary between users) could lead to under-reporting.[36] On the other hand, parental content control tools may actually increase users' propensity to report content, even if the child experiences fewer instances.[37] Relying solely on metrics based on user reporting to detect harmful content may not be a reliable way to assess the impact of a safety measure.[38]
- **Attribution issues due to non-service changes:** We should account for external factors (outside of the service) that could affect content on the service (e.g. a spike in harm level due to a terrorist incident or war) to avoid drawing incorrect conclusions about effectiveness of the measure.
- **Attribution issues due to platform changes:** Platforms are constantly innovating their services and safety measures and may make multiple changes at the same time that could affect content shown on a particular surface (e.g. in a recommendation feed). Therefore, it may be difficult to isolate the effect of changes from a specific safety measure.
- **Inaccurate base of users**: We note the difficulties linked with measuring the number of child accounts, particularly where children have been able to circumvent age restrictions.[39] This limits the robustness of interpreting metrics where the number of known child accounts on the platform does not reflect the actual number of child users.
- **Perceived harmful content:** One measure of how successful parental content controls have been relates to how well parent choices (i.e. through keyword filters) are reflected in what their child sees, which relies, in part, on the

[36] For example, a vulnerable person actively searching for adjacent or pro-eating disorder content may not report this content as unsafe. This may also be the case for underage users, who will not all possess the self-awareness and levels of risk aversion to harmful content and may not understand or report content that is dangerous for them or content that should be generally restricted.

[37] A child account reporting more restricted content may suggest they are exposed to more restricted content. However, wider aspects of parental tools such mediation and media literacy also aim to help child users recognise dangers on the service, and therefore increased reporting may indicate that the child is more aware about what content is restricted and shouldn't be available to them.

[38] Ofcom research suggests for example that some users may not report suspected fraud for reasons such as embarrassment at being a victim of fraud: Ofcom prepared by Yonder Consulting, 16 March 2023. Executive Summary Report: Online Scams & Fraud Research. https://www.ofcom.org.uk/__data/assets/pdf_file/0025/255409/online-scams-and-fraud-summary-report.pdf [Accessed on 31 January 2024].

[39] Ofcom's Children and parents: media use and attitudes report showed that 71% of 8-11-year-olds have their own profile on at least one of the social media sites listed in our study:
Ofcom, 29 March 2023. Children and parents: media use and attitudes report 2023. https://www.ofcom.org.uk/research-and-data/media-literacy-research/childrens/children-and-parents-media-use-and-attitudes-report-2023 [Accessed on 31 January 2024].

effectiveness of a platform's classification of content. Depending on the design of the parental content control tool,[40] the parental preferences over what constitutes harmful content might vary widely and be specific to the parent and child. There may be challenges to evaluate the effectiveness of systems and processes matched to a very diverse set of individual parental preferences.

User-surveys

3.38 Some useful qualitative metrics and results can come directly from users via user surveys. This can help assess the propensity of users to make use of these tools and their experience of using parental content control tools and any barriers, including ease of use. For example, Ofcom makes use of tracker surveys for VSP and OS to test users' online experiences, which we are refining over time.

3.39 Services may also use consumer surveys, including exit surveys to review the specific reasons why users unpaired their parental content control tools. Awareness of the challenges faced when using user surveys, as well as assessing the qualitative metrics collected, is important. In the context of parental content control tools these include:

- **Low take up:** Parental content control tools currently have relatively low take up, so it is difficult to get robust samples to assess user perceptions and take up of these tools. There might be a lack of awareness about the tools given these are relatively recent interventions. However, if awareness increases, user surveys as a tool could evolve over time.
- **Potential survey biases:** Stated preference bias could lead to a difference between how parents say they use parental content control tools compared to how they actually use them (including social-desirability bias, whereby parents or children tell the interviewer what they believe is the "good" answer rather than the truthful one). Similarly, sample selection bias could mean the sample is biased towards parents who are early adopters of these tools and arguably more "hands-on" parents. These parental characteristics could drive outcomes in terms of their child's online experiences for other reasons (i.e. the parent also employs other ground rules at home, e.g. around "screen time").
- **Measuring perceived harmful content:** When parents put in place settings for their child using the parental content control tool, they may specify particular content they perceive to be harmful. It may be hard to ask a parent to recall the types of content they wanted the tool to control and compare that to the child's experience.

3.40 Underage and vulnerable user issues: There are several challenges that come with surveying children to understand their online experiences. For example, there are ethical considerations about asking children about the harms they encountered. Further, it may be hard to get accurate responses, as child users may not understand what constitutes harmful material and may tell a surveyor what they expect they want to hear, leading to underreporting of harms.

3.41 Some ways to mitigate these effects include using large samples or more targeted surveys to overcome take-up issues. Larger samples can allow for more detailed follow-up questions to test users on reasons behind different choices. Collecting data from a number of sources,

[40] In some cases, the parental control controls that offer users to filter out content matching keywords may provide a service curated default list of keywords to select from rather than the parent inputting their own keywords or terms.

not only from user responses to surveys, may also allow for further cross-checks either to corroborate results or estimate the possible under/over-statement in surveys. The difficulty with such solutions is that they can add complexity and expense.

Wider evaluation approaches

3.42 To mitigate some of the challenges raised above, a range of methodologies could be considered, while keeping the evaluation efforts proportionate to the potential risks and harms that the safety measure seeks to address.

3.43 For example, more qualitative 'behavioural audits' of the choice architecture and design of the parental supervision tools on the service could help identify reasons behind apparent lack of take-up, ease of use and engagement with tools. Methods such as field experiments, in a real-world setting, can also be used to test user responses to different implementations of a tool, although they require sufficient numbers of users using the tool in different configurations to produce robust results. As noted in Part I, Ofcom has also used online randomised control trials to test likely responses to different configurations of safety measures.[41] However, RCTs are only suitable when it is possible to have a control group (in this case parents who are not offered the content controls) but this would raise ethical considerations.

3.44 In addition, econometrics studies discussed in Part I can be used to control for external/confounding factors that might influence final outcomes. This could be particularly useful for example to test harm outcomes across accounts and services with parent content control in place (and allowing for different configurations) relative to those without parental content control tools. This would typically rely on large dataset and well-specified models to appropriately control for those confounding factors.

Step 4. Evaluate and Iterate

3.45 Online safety measures like parental content controls are relatively new, making ongoing impact assessments critical as a first step in getting the appropriate systems and processes in place. Ongoing evaluation is likely to provide valuable insights as to their efficacy and the role they could play in delivering online safety.

3.46 An evaluation of specific safety measures would allow for an understanding of what works and how it leads to observed outcomes. Having worked through the previous stages when applying this framework, services should better understand whether the introduction of parental content control tools has resulted in the intended outcomes. Based on the outcome of the evaluation, it would also allow for refinements to be made and highlight whether additional safety measures need to be in place to secure the intended protections.

[41] See for example in Ofcom, 20 June 2023. Nudging users to report potentially harmful online content. https://www.ofcom.org.uk/news-centre/2023/nudging-users-to-report-potentially-harmful-online-content [Accessed on 31 January 2024].

4. Concluding remarks and next steps

4.1 Online safety issues are rapidly evolving as online threats and technologies change. Evaluation will be a critical pillar of online safety to assess the effectiveness of safety measures, to highlight best practice and encourage continuous improvement. This discussion paper sets out an example approach to evaluation in online safety and we welcome stakeholder views and challenge on our thinking.

Next steps

4.2 We intend to engage with experts and industry to understand different approaches to evaluating the effectiveness of safety measures used on online services. Given the novelty of online safety regulation, evaluation will need to evolve and iterate in light of lessons learnt.

4.3 Global efforts to improve online safety include, among others, the EU which recently launched the Digital Services Act with data and reporting requirements, and Australia's online regulator (eSafety) which is also seeking to evaluate safety measures. Furthermore, there have been some global efforts via the World Economic Forum and existing initiatives from some services to coordinate over transparency about their safety approaches.

4.4 Ofcom will continue to examine and test approaches to evaluation of online safety measures, as well as leverage global thinking and expertise and coordinate approaches with international partners where appropriate.

4.5 We welcome expressions of interest from service providers who would be willing to collaborate on testing and trialling different safety measures in a real-world setting. See Annex 1 for information on how to contact us about this paper or discuss collaboration opportunities.

A1. Responding to this Economic Discussion Paper

How to respond

If you would like to respond to this Economic Discussion paper, you can reply using any of these options:

You can respond by email to edp.responses@ofcom.org.uk. If your response is a large file, or has supporting charts, tables or other data, please email it to edp.responses@ofcom.org.uk, as an attachment in Microsoft Word format, together with the cover sheet.

Responses may alternatively be posted to the address below, marked with the title of the EDP:

Economics and Analytics Group
Ofcom
Riverside House
2A Southwark Bridge Road
London SE1 9HA

We welcome responses in formats other than print, for example an audio recording or a British Sign Language video. To respond in BSL:

- send us a recording of you signing your response. This should be no longer than 5 minutes. Suitable file formats are DVDs, wmv or QuickTime files; or
- upload a video of you signing your response directly to YouTube (or another hosting site) and send us the link.

We do not need a paper copy of your response as well as an electronic version. We will acknowledge receipt of a response submitted to us by email.

A2. VSP and Online Safety duties for relevant online services

Since 2020, Ofcom has been the regulator for VSPs that fall under UK jurisdiction.[42] Following the OSA, which received Royal Assent on 26 October 2023, Ofcom is now responsible for regulating a wider set of services that include search services and user-to-user online services in the UK.[43]

The VSP regulation continues to apply to all in-scope VSPs for a transitional period that commenced on 10 January 2024. During the transitional period most OSA duties will not apply to VSPs in scope of the VSP regulation. The transitional period will end, and the VSP regime will be repealed in full, at a date to be set out in secondary legislation by the Secretary of State. Once the transitional period ends, VSPs will be fully regulated under the OSA.[44]

While there are some differences between the regimes, both the VSP regime and the OSA have as their overarching objective to keep users safe from illegal content and to protect children from content that may harm them.

VSP regulation

Under the VSP regime, VSP providers must determine which of the safety measures listed in Schedule 15A of the legislation it is appropriate for them to take to protect all users on their platform from videos containing 'relevant harmful material' and under-18s from videos containing 'restricted material'. VSP providers must determine whether it is appropriate to take a particular measure based on whether it is practicable and proportionate to implement that measure, considering factors such as: the nature of the material on the platform and the potential harm it may cause; the characteristics of users of the service (for example, under-18s); the size and nature of the service; the rights and legitimate interests of users, service providers, and the general public. The chosen measures must be implemented in a way that is effective at protecting users from harmful material.

OSA Regulation

The OSA imposes duties on relevant services to take appropriate steps to keep their users safe from risks associated with illegal content and to provide additional protections for children from harmful content. The OSA captures a wider set of online service types (user-to-user services not only VSPs, search, and sites hosting adult content) with slightly different rules for each.

---

[42] The VSP Regime is set out in Part 4B of the Communications Act 2003 (the Act) and derives from the European Audio-visual Media Services Directive (AVMSD) 2018. The requirements for platforms came into effect in November 2020. For further information, see Ofcom, 11 January 2023. Notified video-sharing platforms. https://www.ofcom.org.uk/online-safety/information-for-industry/vsp-regulation/notified-video-sharing-platforms [Accessed 31 January 2024].

[43] See Ofcom, 26 October 2023. Online safety rules: what you need to know. https://www.ofcom.org.uk/online-safety/advice-for-consumers/online-safety-rules [Accessed 31 January 2024].

[44] For more information on the transitional period, see Ofcom's guide to the repeal of the VSP regime.

Similar to the VSP regime, the OSA recognises that appropriate and proportionate measures will vary depending on the size and risks of the service. The regime places duties on relevant services to conduct risk assessments and children's access assessments. Based on those assessments, services then need to put in place proportionate safety measures, systems and processes to mitigate and manage risks.

Ofcom is producing Codes of Practice, that will set out safety measures services could or should take to demonstrate compliance. Ofcom has published consultations, including setting out its draft Codes of Practice in relation to illegal harms and most recently, child protection.

A3. Potential factors in prioritising metrics

A3.1 Below is a list of factors that could be used to prioritise metrics:

Table A1: Possible prioritisation factors

Main factors	Example considerations
Salience for measuring effectiveness	Relation to the objective: intended outcomes should be prioritised over potential unintended outcomes. Implementation and outcomes: in general, we want to understand whether something has happened (e.g. a safety measure was implemented and lead to expected outcomes), as a priority over the steps in between (how or why something has happened), unless we need metrics to dig deeper.
Ease of analysis and robustness	Objectivity: simple and numerical metrics are preferred to those that are implied and subjective. Standalone: metrics where there is inherent value in the metric are prioritised, as opposed to those which require additional information and context to be interpreted meaningfully. Comparable & consistent: metrics that are relevant across services and are likely to remain relevant in the future are prioritised. Historical data: we would prioritise metrics where historical data is available. This is so that we can assess trends over time in light of regulation and/or before and after introduction of a safety measure. Alternative metrics: metrics where few alternatives exist to capture the same process/outcome are prioritised. However, in some cases more than one metric might be needed if there are different biases and sources of error relying on a single metric to measure a process, risk or outcome.
Practical, cost effective, and appropriate	Cost considerations: metrics where the data collection is less costly for Ofcom and/or for providers are preferred. Ease of access: metrics from third parties and data suppliers that will engage more quickly are prioritised, as are metrics from Ofcom's own consumer research. Privacy & ethics: metrics that do not pose privacy or ethical risks are prioritised.

A4. Keywords list

A4.1 While the main audience for this paper are Trust & Safety experts and the academic community working on online safety issues, Box 3 below provides descriptions of key terms related to online safety for those less familiar with the terminology or concepts.

Box A1: Key terminology and concepts in online safety

Age assurance measures: refers to the range of measures that can be taken by an online service provider to be informed about a user's true age.

Hash-matching: is a term to describe the technical method in which one image is algorithmically matched to another. Hash matching can be used in the detection of illegal or harmful images or videos.

Moderation: refers to the methods services use to monitor and restrict content on their service which may violate their terms & conditions or community guidelines. Moderation can be proactive, in that the service takes steps to check and detect content sometimes even prior to publication often using automated tools, or reactive in response to content that has been reported.

Platforms and services: are digital spaces on which users can consume and/or create content. These may be VSPs (under the VSP regime) or search and user-to-user services (under the broader Online Safety regime). In this report we use 'services' to refer to both as a collective, and 'platform' only when it is specific to VSPs.

Restricted material: relates to content that would receive an R18 classification or likely to be refused classification by the BBFC as well as material that that might impair the physical, mental or moral development of under-18s.

Search services: means an internet service that is, or includes, a search engine allowing users to search for specific types of content.

Surface: refers to a specific area that a user sees within a service. For example the "for you" page or the "friends" page.

Terms and conditions: mean any rules, policies or standards communicated to users governing the type of actors, behaviour or content permitted or not permitted on an online service, including, but not limited to, community guidelines, community standards, terms of service or specific content-related policies.

User empowerment tools: are measures that allow users to take a greater degree of control over their online experience particularly the type of content that they see on the service.

User policies: include a service's terms & conditions and community guidelines. They are a set of rules regarding how a user should act on an online service.

User(s): means anyone able to access, view or upload videos on an online service such as a VSPs, not just those who have an account. In the context of certain protection measures, a user could also be a parent or guardian.

User to user service: means an internet service where content is generated directly on the service by a user of the service, or uploaded to or shared on the service by a user of the service, which may be encountered by another user, or other users, of the service.

A5. Data sources

A5.1 In this annex, we discuss some of the data sources we might use and some consideration around their use.

Table A2: Potential data sources to be used in evaluation

Data sources	Description
Requested information from services	We have powers to request data from services via formal information requests and, under the Online Safety regime, transparency notices information requests. We can request existing data, research and user testing information that services hold, but we can also request services begin to collect new data. Services may also capture data via KPIs provided by third-party providers, i.e. where they outsource moderation.[45]
Service APIs, desk research on publicly available data and web scraping	We have been able to collect some data from services' sites based on our own desk research or using APIs[46] and web scraping tools. These latter tools can automate and streamline information gathering processes, but to date have required permission of the service where their use is not expressly permitted for regulatory purposes. Given the relative efficiency of this collection approach, we expect to continue to discuss with services using this method where appropriate, while recognising issues and potential limitations around processing and recording of personal data, copyright and service's terms more generally.
Third party datasets	Commercial third parties collect industry data, particularly user data on usage and usage journeys on different services (reach, time spent on services, prior or subsequent services visited etc.).[47] Current data is more limited around service's internal processes.
User surveys	Services often conduct surveys for example to understand preferences and reactions to product changes and could be extended to test online harm experiences). Ofcom has an extensive ongoing research programme, e.g. Online Experiences and VSP Trackers on attitudes and experiences. Some survey evidence may also be available from other regulators[48], third-sector organisations.[49]

[45]Telus International reports on key operational metrics, such as Content Moderation Accuracy, Average Handling Time, Attrition rate, see: Content moderation solutions.

[46] APIs are made available for some platforms and provide a means to gather large datasets that researchers and regulators can use to conduct analysis.

[47] There are a range of third-party providers with different specialisms, including Ahrefs, Apptopia, Crunchbase, SimilarWeb.

[48] See Online Safety in Australia- Adult's experiences online.

[49]See Substack, 22 June 2023. Designing Tomorrow. Unveiling the Neely Ethic & Technology Indices. https://psychoftech.substack.com/p/unveiling-the-neely-ethics-and-technology [Accessed on 31 January 2024].

Data sources	Description
Government organisations and agencies	For the most harmful material, official statistics and law enforcement are also likely to be important avenues for aggregated metrics. For example, the Office for National Statistics (ONS) collects police recorded crime data, and the national CSEW (Crime Survey of England and Wales).
Third-sector specialists	A wide range of organisations collect data on individual harm areas, particularly on CSEA. Third-sector organisations representing children or users with protected characteristics may be better able to give a voice to some user communities and their experiences.
Complaints data	Ofcom also has a complaints tool for users to raise concerns about services. Via our internal triage function, we collate information by service and complaint types. Such data is intended to enable users to raise concerns about online safety but can also provide further data on any emerging risks or harms for UK users.