Ofcom
making communications work
for everyone

Behavioural insights for online safety: understanding the impact of video sharing platform (VSP) design on user behaviour

Economic discussion paper series, issue number 4

Re-issued 14th July 2023

Originally published 20th July 2022

Notice of revision

We originally published this research in July 2022. This paper contains updated analysis to that research for the alert messages trial. In completing our most recent online trials work, an inconsistency was identified in the data capture, and therefore analysis, between the alerts messages trial and subsequent Ofcom trials**.** Working with Ofcom, Kantar Public suggested a revised method which would sufficiently rectify the issue and make the analysis comparable across trials. With Ofcom's agreement, Kantar Public completed the re-analysis set out in the revised trial report.

The main change in the results is that only the high-level descriptive social proof alert message had a statistically significant impact on participants' skipping behaviour, whereas in the previous analysis all interventions had a statistically significant impact on participants' behaviour. We have updated the presentation of our results in the overview and section four to reflect this change.

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 inform that debate. 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.

Acknowledgements

Ofcom would like to recognise the assistance provided by Kantar Public's Behavioural Practice who we commissioned to provide the experimental design and set-up of the online trials and for the detailed analyses set out in Appendices 1 and 2.

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.

Contents

1. Overview

1.1 Most internet users believe the benefits of going online outweigh the risks.[1] However, seven out of ten report having seen or experienced something potentially harmful in the last three months.[2]

1.2 Online platforms commonly use safety measures such as alert messages, reporting mechanisms, parental controls and terms and conditions to safeguard their users. However, there is limited research in the public domain about their effectiveness. At the same time, there is growing behavioural insight evidence that relatively small changes to the design of the online environment, which would include safety features, can shape user behaviour - for better and for worse.[3]

1.3 Ofcom has a duty to promote media literacy, including in respect of material available on the internet. Ofcom's approach to media literacy is multi-dimensional and considers a range of aspects including how the design of services can impact on users' ability to participate fully and safely online.

1.4 In this Economic Discussion Paper, we discuss results from research into the impact of behaviourally-informed designs of **alert messages** and **content-reporting** mechanisms on video-sharing platforms (VSPs) on user behaviour. Alert messages are used by VSPs to warn users about the potential harmfulness of the content they are about to watch. Content-reporting mechanisms allow users to report potentially harmful content to VSPs. Our research is intended to improve our understanding of the effectiveness of these safety measures. We explore how the design of alert messages and content-reporting mechanisms affects the decisions people make about viewing potentially harmful content and reporting such content.

1.5 We chose these safety measures to research because: they are widely used by VSPs; there is limited research in the public domain on their effectiveness at improving safety; and we were able to replicate users' interactions with these safety measures in an experimental setting.[4]

1.6 This research is Ofcom's first use of an online randomised control trial ('RCT'). In an RCT some research participants are randomly allocated to a 'control' group and their behaviour is compared to the behaviour of others randomly allocated to 'treatment' groups. In our online RCT all participants interacted with a mock-up of a VSP interface and had the opportunity to view a series of short videos: some containng 'neutral' content and some

[1] Ofcom, October 2021, Video Sharing Platforms: Ofcom's Plan and Approach.
[2] Ofcom, October 2021, Video Sharing Platforms: Ofcom's Plan and Approach.
[3] CMA, April 2022, Online Choice Architecture – How digital design can harm competition and consumers.
[4] Other interventions we considered, such as uploading user-generated content, are more difficult to replicate in an experimental setting.

containing legal but potentially harmful content.[5] In our treatment groups, we vary our platform's user interface compared with a 'standard' interface in the control group (for example by making the reporting button more salient in the treatment group). By randomising the allocation of participants between the control and treatment groups, we are able to isolate the difference in behaviour caused by the variations in the user interface. This allows us to draw conclusions about the causal impact of the variations.

What we have found

Alert messages

We found that alert messages featuring a high-level descriptive social proof message could be an effective mechanism to help users to make more informed choices about legal but potentially harmful content.[6]

In the control group, research participants did not encounter any alert messages prior to viewing videos. We then looked at how the likelihood of skipping the legal but harmful videos changed when participants were exposed to one of the following alert messages that appeared before they watched a video containing legal but potentially harmful content:

1. A generic warning message: "This video may contain sensitive material".
2. A high level descriptive 'social proof' message indicating how other users have reacted to the video: "This video contains material that *other viewers on this platform have reported* as being sensitive".[7]
3. A specific warning about the nature of the video the user was about to view, e.g., "This video contains misinformation".

The high-level descriptive social proof alert message led to a statistically significant increase in the likelihood of trial participants skipping legal but potentially harmful content compared to the control arm. The other two alert messages also increased skipping but not to our threshold level for statistical significance.[8] Although there was no statistically significant difference in the effectiveness of the specific content warning alert messages on overall skipping behaviour, our research suggests that the specific content warning message was more effective at prompting participants to skip at the alert message stage (rather than after the video had started).

We also found that introducing alert messages did not increase the likelihood of skipping neutral content. That is, the alerts did not prime trial participants to skip other types of content as well.

---

[5] Ofcom did not designate the videos as legal but potentially harmful. This task was undertaken by Kantar Public, the research agency we commissioned to undertake the research.

[6] There is not yet a formal definition of 'legal but harmful content'. In the context of the Online Safety Bill, the government has indicated that it will set out in secondary legislation a number of priority categories for "legal but harmful" content. See Online Safety Bill Factsheet. Accessed on 29th June 2022.

[7] Italics were not used in the text seen by research participants.

[8] The threshold used for statistical significance was set at $p<0.05$.

In addition, the majority of research participants did not find the alert messages to be annoying and those who were exposed to alert messages and chose to watch the video content did not regret their decision to do so. The research participants who were exposed to an alert message but who then continued to watch a video, on average, watched the legal but potentially harmful content for longer than participants in the control group. The probability of them reporting this content also reduced. Finally, research participants reported that they found the specific content warning alert message to be more useful than the generic warning alert message.

Taken together, these findings suggest that alert messages prompted research participants to become more engaged in making decisions about watching the video content they were presented with in the course of the experiment but only the high level descriptive social proof alert message had a statistically significant impact on whether participants skipped legal but potentially harmful content. This form of alert messages gave those participants who did not wish to view the content the information to enable them to decide to skip that content. Equally, those participants who chose to view the content after being exposed to an alert message were better placed to understand the risk of harm: they were either comfortable with the content or were alive to the fact that they might find the content problematic and so were 'primed' to skip away from it if necessary.

Reporting mechanisms

Our experiment on the impact of reporting mechanisms suggests that making the reporting function more prominent encourages users to report content that they are concerned about.

Participants in the control group used the mocked-up VSP interface with the reporting function as an option behind a button marked with an ellipsis ('…'). We tested three treatments in respect of reporting mechanisms:

- Raising the prominence of the reporting option on the user interface using a 'flag' icon. This made the option more salient to participants.
- Raising the prominence of the reporting option *and* adding a prompt to ask them if they would like to report content if they had disliked or commented on a video. We called this the 'salience plus prompt' treatment.
- Using the salience plus prompt approach from the second treatment and simplifying the reporting process.

All three treatments increased the likelihood that the legal but potentially harmful videos would be reported by the research participants.

In our experiment, the level of unprompted reporting of legal but potentially harmful content in the control arm was extremely low: 98% of research participants in the control arm did not report any of the legal but potentially harmful videos even though around 23-27% of research participants chose to 'dislike' that content.

Raising the prominence of the reporting option on its own significantly increased the likelihood that the legal but potentially harmful content would be reported. For instance, in the control group only around 1% of research participants reported one of the legal but potentially harmful

videos. Raising the prominence of the reporting function increased this to 4% - a fourfold increase.

Raising the profile of the reporting function *combined* with an additional prompt to report for participants who disliked or commented on legal but potentially harmful content had a more significant effect on increasing the likelihood that the content would be reported. For instance, with this intervention 11% of research participants reported one of the legal but potentially harmful videos compared to 1% in the control group. Furthermore, 4% of research participants reported all three legal but potentially harmful videos compared to 1% in the control group. This result could suggest that participants had been using the dislike or comment function as a proxy for reporting content but that a targeted prompt could encourage them to use the formal reporting process.

We did not find any significant difference between raising the profile of the reporting function together with an additional prompt, on the one hand, and combining this intervention with a simplification of the reporting process on the likelihood of reporting, on the other. We did not find that participants struggled to complete reports once they had decided to report content so it did not appear that simplifying the reporting process had any incremental impact on the volume of reporting.

We also tested if there were unintended consequences from the changes we made to the user interface. For example, whether the accuracy of reporting declined as the likelihood of reporting increased or whether raising the profile of the reporting function could lead participants to 'over-report' content, even if that content is not potentially harmful. We found that in the salience plus prompt and the salience plus prompt plus simplification treatment arms, there was a statistically significant increase in the likelihood of number of legal but potentially harmful videos being accurately reported compared to the control arm. That is, research participants not only reported more legal but potentially harmful videos but they also categorised the videos correctly.

The levels of the over-reporting of neutral content across both the control and treatment arms were so low such that it was not possible to make any reliable inferences about whether the changes to the interface led to any change in the volume of over-reporting.

Taking these findings together, our research on reporting mechanisms suggests that increasing the salience of the reporting function can increase user reporting of legal but potentially harmful content and does not necessarily lead to a reduction in accuracy or an excessive reporting of neutral content.

1.7 There are benefits and limitations to using an online randomised control trial ('RCT') with a mock-up user interface. The main benefits are that the user experience is similar to their actual interactions with platforms and that we can make use of a relatively large sample (in

this case approximately 2,400[9] participants in each of our two experiments). This increases our confidence that our results are statistically significant rather than due to chance.

1.8 The main limitation of our research is that it is generated in an experimental environment. Participants are aware they are taking part in research and this can potentially distort their behaviour.[10] And while we were able to create a realistic platform and saw reassuring signs of participants interacting with the user interface in familiar ways (such as unprompted 'liking' and commenting on content), it was not possible to replicate the wider context in which consumers interact with VSPs. In addition, we were not able to measure the impact of ongoing exposure to variations in the choice architecture of the safety measures. This matters because a large amount of online behaviour is repeated on a daily basis over extended periods of time. The experimental set-up allows us to assess an intervention at a point in time but it does not allow us to observe the medium- and long-term effectiveness of the interventions we tested.

1.9 As a result, there are limitations on how far we can extrapolate from behaviour identified in this simulated setting. Nevertheless, online RCTs provide useful insights. They have been adopted as evidence-building tools by other UK regulators too and we will continue to make use of this methodological approach.[11] Ideally, future research could extend to running RCTs in collaboration with online platforms so that we can test changes to safety features in real-world settings. We would welcome the opportunity to explore the scope for conducting RCTS in collaboration with industry stakeholders.

1.10 We hope that our experiments will prompt further debate and research on these issues.

1.11 This paper is structured as follows:

a) Section two describes the context for the research.

b) Section three describes our research approach.

c) Section four sets out our results and insights and discusses the limitations of our research.

d) Annex 1 contains our review of the evidence on users' behaviours in respect of alert messages and reporting mechanisms.

[9] In the experiment involving alert messages, there were 2,401 research participants. In the experiment involving reporting mechanisms, there were 2,400 research participants.

[10] The Hawthorne effect is a term used to describe the phenomenon that some participants in experiments modify their behaviour as a result of being observed. For instance, see Adair, J.G. (1984) The Hawthorne effect: A reconsideration of the methodological artifact, *Journal of Applied Psychology*, 69(2).

[11] See, for example, FCA Occasional Paper 51 Using online experiments for behaviourally informed consumer policy.

2. Research context: protecting users online

2.1 Video sharing platforms like TikTok, Snapchat and Twitch are a huge part of online life. These platforms offer many benefits – keeping us entertained and informed on big issues and offering a platform for creativity and self-expression. For instance, we use them for entertainment; to engage with friends and people we know; and to take part in debating issues of the day with people from a range of different perspectives.

2.2 However, being online is not without risk. Ofcom research has found that seven in ten users of these platforms have seen or experienced something potentially harmful.[12] A third of people witnessed or experienced hateful content, and one in five saw videos or content that encouraged racism.[13]

2.3 Ofcom is using behavioural insights to help understand how users behave in real life and, in this context, how effective the safety measures that online platforms put in place are.

2.4 Traditional economics assumes that people are always rational, make decisions based on their own self-interest, and will change their thinking as new information emerges. In the rational model, people weigh the costs and benefits of action, considering all the available information and options. Whilst this approach can be a reasonable approximation to people's behaviour in some contexts, it is not always the way that people make decisions.

2.5 Behavioural economics takes a different approach by assuming that people make decisions in an imperfect but predictable way.[14] In behavioural economics, people have cognitive limitations (they are unable to take into account all the information available to them and make use of mental short-cuts or heuristics) and are subject to cognitive biases.[15] That is, their preferences are not always fully formed and their decisions can be shaped by the decision-making environment. For example, decision-making can be influenced by the starting point for their choices, how choices are presented to them, and how choices are described. The term 'choice architecture' is used to describe the contexts in which people make decisions and how choices are presented.[16]

2.6 The cognitive factors affecting people's decisions can be exacerbated in an online environment. This may be because there is more information online than people can process, they can face more immediate decisions than they do offline, or they may make decisions on their own without being able to confer with a friend or family member. In addition, online platforms have considerable control over which choices are presented to

[12] Ofcom, October 2021, Video Sharing Platforms: Ofcom's Plan and Approach.
[13] Ofcom, October 2021, Video Sharing Platforms: Ofcom's Plan and Approach.
[14] Ariely, D. (2009), *Predictably Irrational: The Hidden Forces that Shape Our Decisions*, HarperCollins.
[15] Kahneman, D. (2011), *Thinking Fast and Slow*, Farrar, Straus and Giroux.
[16] Thaler, R. H., Sunstein, C. R., and Balz, J. P. (2010). Choice Architecture.

people and how those choices are presented, even integrating choice architecture in their algorithms.[17]

2.7 Platforms' user interfaces can be designed to help promote users' interests. For example, platforms can mitigate people's cognitive limitations and biases by providing information in a clear and balanced way. On the other hand, a platform's user interface can be designed (either deliberately or inadvertently) to obstruct users from exercising choices in their own interest e.g., through making information more complex than it needs to be or by making it more difficult for the user to make a particular choice.[18]

2.8 Our research is intended to contribute to the understanding of the effectiveness of the safety measures we have researched. It can also serve to stimulate a dialogue with industry stakeholders on both the effectiveness of those safety measures and also on the different methods that are available for researching their effectiveness. There will be a need to keep our research under review as safety measures and users' expectations of protection against online harms evolve over time.

[17] CMA, January 2021, Algorithms: How they can reduce competition and harm consumers.
[18] CMA, April 2022, Online Choice Architecture – How digital design can harm competition and consumers.

3. Research approach: designing experiments to test the effectiveness of online safety measures

Introduction

3.1 In this section we outline how we selected the safety measures to research and how we selected the interventions to test for each safety measure. We also explain our experimental set-up and how we addressed ethical considerations in that experimental set-up.

Choosing safety measures to test

3.2 We considered the range of safety measures used by platforms as potential candidates to research.[19] We prioritised our choice of safety measures along the following criteria:

a) availability of research on the effectiveness of the safety measure and whether Ofcom's research could contribute to the evidence base;

b) evidence that there could be barriers to users engaging with the safety measure; and

c) the likelihood that behavioural research into the safety measure using our experimental set-up would yield useful results.

3.3 Using these criteria, we decided to research alert messages and reporting mechanisms in relation to legal but potentially harmful content.

Research Methodology

3.4 The research methodology was an online RCT, Ofcom's first use of this research approach.

3.5 In an RCT, research participants are randomly assigned to either a 'control' group or a 'treatment' group. Randomising users across control and treatment groups removes selection effects. Selection effects are a bias introduced to research when the sample in the research is biased towards a subset of the population the researcher intends to assess, because some people might self-select into particular groups. Using randomisation coupled with control and treatment groups allows us to directly compare the impacts of the changes to the user interface on people's behaviour and to assess causality.[20]

[19] The safety measures we considered included: age verification, alert messages, parental control measures, reporting mechanisms, and terms and conditions.

[20] These features mean that RCTs are often described as the 'gold standard' for studying causal relationships and evaluating the effectiveness of interventions. For example, Hariton, E. and Locascio, J. (2018) Randomised controlled trials: the gold standard for effectiveness research,

3.6 For our online RCTs we used a mocked-up VSP interface which was similar to a real-world VSP interface so that the user experience could be close to the actual experience of using a VSP. Our online RCTs should thus have greater ecological validity than a laboratory-based RCT.

3.7 Using an online RCT also means that we can make use of a relatively large sample. We had approximately 2,400 participants in each experiment compared to perhaps a few hundred in a typical lab-based setting.[21] This increases our confidence that our results are statistically significant rather than due to chance.

Choosing interventions to test

3.8 In this section we describe the interventions we chose to test and our outcome measures. We summarise the evidence from the behavioural literature relevant to the safety measure and our observational research which helped us to decide which interventions to include in our research. Further details of that evidence are set out in Annex 1.

Alert messages

3.9 The purpose of alert messages for legal but potentially harmful content is not to prevent people from viewing content. Instead, alert messages help people to make a more informed choice about whether to watch the video, given the information in the alert. We assume that, in general, platforms will want to disrupt user engagement with their service as little as possible. Equally, users will want interruptions to a service to be clearly justified and easy to deal with.[22] We thus assume that both parties will want alerts that are parsimonious i.e., the least disruptive mechanism that provides users with informed choice.

Behavioural literature

3.10 In general, alert messages work by introducing frictions to make users pause to consider whether to watch a video.[23] But the strength and design of those frictions can vary substantially. The behavioural literature points to a number of conceptual aspects of choice architecture that are thought to influence their effectiveness:

3.11 **Timeliness**: a central tenet of behavioural insight is the increased potential to shape behaviour if prompts are provided close to or at the moment of choice.[24] Warnings can be placed when users sign up to a service (as is the case with terms and conditions), at the

[21] In the experiment involving alert messages, there were 2,401 research participants. In the experiment involving reporting mechanisms, there were 2,400 research participants.
[22] For instance, see GDS Design Principles guidance.
[23] This is the difference between thinking 'fast' using heuristics and defaults and thinking 'slow' by being more deliberative. See Kahneman, D. (2012), *Thinking Fast and Slow*, Penguin.
[24] Behavioural Insights Team (2014), EAST: Four simple ways to apply behavioural insights.

point of consumption (alerts on content as the user first engages with it), or even during content consumption (in-clip warnings about specific components of video clips).

3.12 **Salience**: our attention is drawn to information that is novel and seems relevant to us.[25] Designers of warnings can vary the salience of the warnings in multiple ways. The size of warnings (e.g., full screen versus partial screen), the colour and font of the text, whether the text is static or moving, and whether warnings are verbal or in the form of images. The evolution of warnings on cigarette packaging provides a useful example: developing from written warnings in relatively small font sizes to whole pack warnings with graphic images of harm.

3.13 **Friction and defaults**: some of the defining developments in social media design, such as autoplay and 'infinite' scroll are centred on removing frictions - allowing users to consume content without having to take any actions at all or removing frictions that can make them stop. With alerts, the default setting that is linked to the alert will have a significant bearing on user behaviour. For example, a default to autoplay content with an alert, unless the user actively skips forward, is likely to cause a significant difference in viewing compared to a default to skip content carrying an alert, unless the user actively clicks to watch the content.

3.14 Equally important are the factors that will influence decision-making on the **user side**. Several studies have established that we use heuristics, or short cuts, and look for clues in our environment to help us make decisions.[26] This could well be the case when browsing the internet when we are frequently scanning material and making rapid judgements on what to watch and what to skip. Factors that shape these decisions include:

3.15 **Trust and credibility**: whether we heed a warning will depend on whether we think the warning is credible – and that credibility will depend on a number of factors, ranging from whether we have encountered similar warnings in the past that we have found helpful, to whether we trust the source of the warning (e.g. whether the source is a platform or other users) and whether others seem to be heeding the warning (e.g. a warning on content that has a large numbers of likes may seems less credible than one with multiple dislikes).

3.16 **Habit**: in the UK, adult internet users spend almost four hours online per day.[27] As a result, a large amount of user behaviour is habitual. Rather than making individual decisions about which warning to heed, users may 'automatically' click past pop-up messages out of habit.

3.17 We are aware that the FCA has conducted experimental research looking at the impact of tweeted risk warnings in relation to compliance with financial products.[28] However, there appears to be little evidence in the public domain relating specifically to the impact of

[25] Kahneman and Thaler (2006) Anomalies: Utility Maximisation and Experienced Utility. Journal of Economic Perspectives 20(1):221-234.
[26] Behavioural Insights Team (2019) The behavioural science of online harm and manipulation and what to do about it.
[27] Ofcom 2022, Online nation report
[28] FCA Occasional Paper 47 (2018) The effect of tweet risk warnings on attractiveness, search and understanding.

different types of alert messages or warning messages on users' propensity to view online content.

3.18 There is a small **quantitative** literature on the use of alert messages or prompts in relation to users posting content. Whilst this is quite different from the experience of viewing content, this research can potentially help us to understand relevant cognitive limitations and biases. The research finds that introducing alert messages creates a friction in the user experience and this can (a) prevent people from unintentionally disclosing information, (b) lead users to post fewer offensive posts, or (c) cause users to reconsider sharing content with other people. Overall, this research, whilst limited, indicates that alerts can be effective at encouraging users to consider their choices more fully.

3.19 See Annex 1 for more detail on research related to users posting content.

Observational research

3.20 Our observational research indicated that alert messages are widely used by online platforms and appear in range of different contexts (e.g., warning users of harmful content, warning users that pop-ups have been blocked, etc). A number of platforms use alert messages to warn users about the content they are about to see.[29] The warnings used in alert messages vary considerably across platforms, combining icons and verbal messages. For example, alert messages use terms such as "sensitive", "disturbing", "inappropriate" and "offensive".

3.21 Many platforms also make use of alert messages or prompts to warn users when they are sharing or commenting on content.[30,31] For example:

a) **Commenting on content**: Before a user can post a comment on YouTube, they are reminded "to keep their comments respectful" and that YouTube has Community Guidelines.

b) **Sharing old content**: Facebook provide an alert "This content is more than 3 months old" if a user tries to share content which is more than 90 days old.

c) **Sharing content without reading it**: Twitter provides an alert "Headlines don't tell the full story", if a user tried to re-tweet an article without reading it.

3.22 In some cases, if a user attempts to search for specific types of content on a platform – usually relating to self-harm topics - they receive a prompt which directs them to relevant support services. For instance, Instagram, Snapchat and Twitter all refer the user to relevant local helplines if a user searches for suicide-related content.

3.23 Annex 1 contains some examples of alert messages used by platforms in relation to content that users are about to view.

---

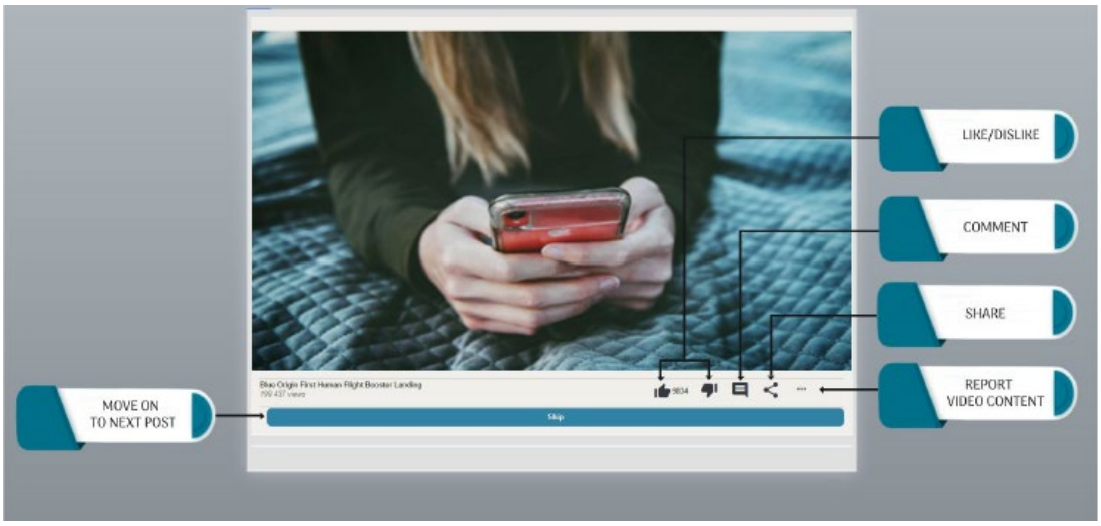
[29] The Verge (2017), Instagram will begin blurring 'sensitive' posts before you can view them.
[30] The Verge (2020), Twitter is bringing its 'read before you retweet' prompt to all users.
[31] The Guardian (2019), Instagram's anti-bullying AI asks users: 'Are you sure you want to post this?'.

Experimental design

3.24 Our experimental setting is a mock-up of a VSP interface. The mock-up includes many of the features typically found on online platforms, such as a 'like' button, ability to comment, and ability to skip videos. The figure below is an example of the user interface in the control – it could be varied according to the intervention we were testing. The interface was device agnostic i.e., it was accessible via a laptop, smartphone or tablet.

3.25 In the experiment, participants were asked to imagine that they were at the end of the day and were watching some videos online. They were then asked to interact with those videos as they would normally. Research participants had the opportunity to watch six videos: three containing 'neutral' content and three containing the content that Kantar Public had considered could be categorized as 'legal but potentially harmful'. Descriptions of this content are set out in Appendices 1 and 2 but in broad terms the three types of content were: Covid-19 misinformation; a fight on a Tube train; and, offensive language. The order in which the videos were shown to research participants was randomized.

Figure 1: Description of user interface in our control group with explanation of icons[32]



3.26 By using a user interface which has similar functionality to the VSPs that people are familiar with when viewing content online, our research participants were able to quickly demonstrate typical online behaviours, such as liking and commenting. The interface also enabled us to alter different aspects of the choice architecture i.e., the way in which choices or decisions were presented to participants in the research.

3.27 In all of the trial arms in which there was an alert message, the clips carrying the alerts had autoplay *switched off*. This meant that users were required to actively decide whether to view or skip. The experiment only tested variations in the phrasing of the alert message.

[32] The descriptions shown in the figure did not appear on the actual interface.

Colour and font were kept the same and no images were used. This allowed us to isolate the impact of variations in the phrasing of alerts.

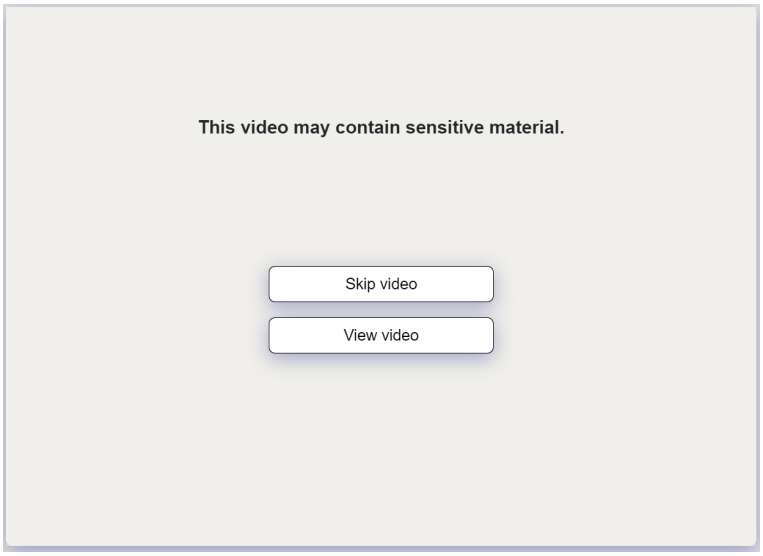
Interventions we tested

3.28 The foregoing information and analysis (behavioural literature, observational research and experimental design) contributed to our assessment of the interventions to test. We prioritized interventions which (a) are used by VSPs, (b) had limited research in the public domain on their effectiveness at improving safety, and (c) could be replicated in our experimental setting.

3.29 This priorisation led us to choose three alerts to test. The three alert messages we tested were: a generic warning; a high-level social proof; and a specific content warning. The control arm did not have an alert message.

3.30 **Generic warning.** A generic warning creates an interruption or friction in the process of moving between videos. This friction gives people the time to consider their actions before deciding to watch the video. The generic warning used language used by some platforms but no behavioural levers (nudges). This warning is the most 'neutral' of the interventions.

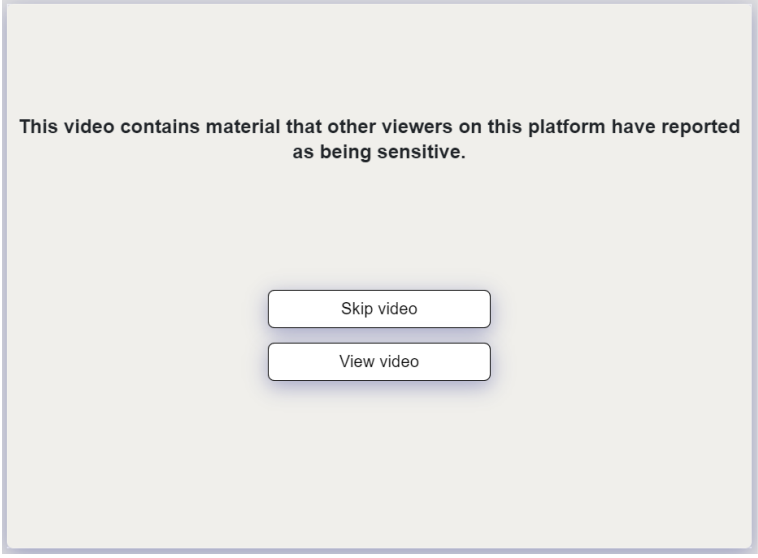
Figure 2: Generic warning message



3.31 **High-level social proof.** In situations of ambiguity, people are thought to be influenced in their decision-making by what others are doing.[33] In a food market in a new city, for example, consumers might look to see which stalls are most popular to decide on where to eat. For this intervention, the alert mobilised social proof by informing the viewer that other users on the platform had reported the content.

[33] Cialdini, R.B. (2007), *Influence: The Psychology of Persuasion*, Harper.

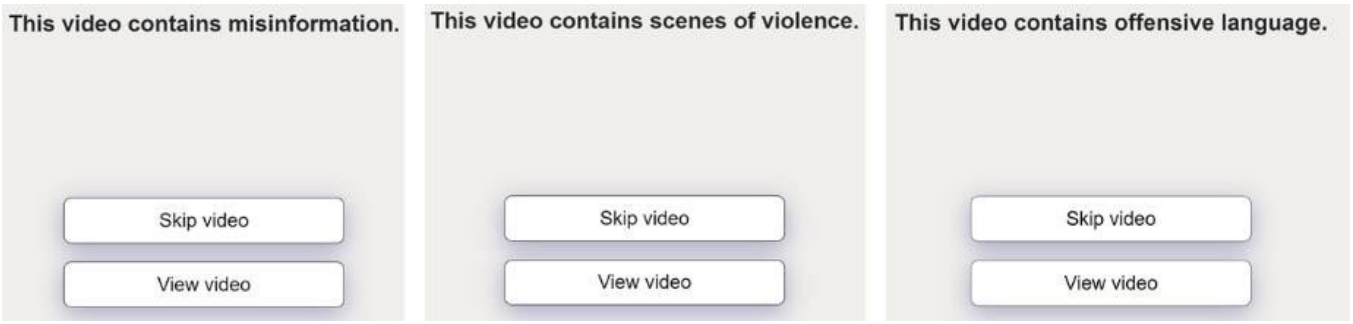
Figure 3: High-level social proof message



3.32 **Specific content warning.** Customising alert messages to the specific content type allows users to make a more informed viewing choice. Whereas social proofing allows users to make judgements on viewing based on the behavoiur of others, a specific warning allows the user to decide on the basis of their tolerance or interest in content.

3.33 Furthermore, the specific content warnings may increase belief in the accuracy in the alert message. For this alert, we used commonly-used categorisations of harm – misinformation; scenes of violence; offensive language – to inform users about the nature of the content in each video.

Figure 4: Specific content warning messages



Primary and secondary outcomes

3.34 The primary outcome is the main change the experiment is designed to analyse. Secondary outcomes are effects which are associated with the primary outcome but are not the focus of the experiment.

3.35 The **primary outcome** of this trial was to understand if alert messages affected the consumption of legal but potentially harmful content (specifically whether users chose to skip the content) and, if so, whether the wording of the alert messages made a difference to users' behaviour.

3.36 The **secondary outcomes** of our research were to:

a) understand if alert messages affected the consumption of neutral content;

b) examine whether there were differences in the viewing time of the legal but potentially harmful content between trial arms;

c) test how comparable the reporting behaviour of research participants was to the reporting behaviour of research participants in our trial of reporting mechanisms (see below); and,

d) investigate research participants' underlying attitudes to the alert messages (for example, whether they found the messages useful or were annoyed by the message).

Reporting mechanisms

3.37 Users have an important role to play in reporting legal but potentially harmful content so that platforms can take steps to review and remove that content (where appropriate).

3.38 However, there may be behavioural barriers which prevent users from starting to report content or only partially complete a reporting process, so that platforms are not able to act on the information. This potentially prevents the reporting mechanisms from being as effective an online safety measure as they could be.

3.39 More effective reporting arrangements also enables online platforms to target their resources on addressing valid concerns or complaints, as well as to improve the personalisation of content shown to users (helping platforms to learn what content users don't like).

3.40 We undertook the same process to identify interventions to test for reporting mechanisms as we did for alert messages: reviewed consumer and academic research, undertook observational research, and developed a list of behavioural barriers that could be relevant.

Behavioural literature and consumer research

3.41 A number of the conceptual aspects of the choice architecture that are relevant to thinking about the effectiveness of alert messages could also be relevant to thinking about the effectiveness of reporting mechanisms. For example, the salience or prominence of the reporting mechanism or the timeliness of prompts in relation to reporting content are factors which could influence the effectiveness of a reporting mechanism. To avoid duplication, we do not repeat those conceptual aspects here.

3.42 We did not find any academic behavioural literature relating directly to the effectiveness of online reporting mechanisms or processes. There is a literature regarding consumer

complaints behaviour. However, this research focuses predominantly on customers' offline complaints about goods and service. It was not clear how applicable this research is to the issue of online content reporting.

3.43 The literature we found concerning online complaints behaviour related primarily to public reviews or complaints, and the effect on the complainant and other consumers. Again, it was not clear how this research was applicable to online content reporting mechanisms. For more information see Annex 1.

3.44 Ofcom's consumer survey research has found that:

a) Although the general awareness of safety measures on VSPs is low, flagging and reporting tools are the most widely known, with 60% of VSP users claiming to be aware of this specific measure.[34]

b) Ofcom's Online Nation report shows that, of the 8% of users who said they experienced their most recent harm on a video sharing site or app, 36% flagged or reported the content. This compares to an average of 31% for potential harms encountered on any type of platform.[35]

c) However, 35% of those exposed to potentially harmful material did not take any action. The reasons for this included a perception that it would not make a difference.[36]

d) Users believe that, once started, they do not find it challenging to complete reports: less than 1% said that they had not reported content because they could not complete the process.

Observational research

3.45 The reporting processes for VSPs and other online platforms are structured in broadly similar ways. That is, once the user has chosen to report content, the platform will typically present the user with a list of broad categories of harm to select from. Once they have selected a broad category of harm, they may be taken to a second page which then gives the user the option to be more specific about the type of harm they want to report. For instance, if the user has selected 'Self-harm' as a broad category of harm, they may then be presented with a sub-set of options such as 'Suicide', 'Intention to self-harm' etc. Finally, they may also be given the option of a free-text box in which they can enter their own description and /or provide additional information before submitting the report. The user typically also has the option of cancelling the complaint at any stage.

3.46 However, there are also potentially important differences in how the reporting function is presented to users. For example, whether it is behind an ellipsis[37] or represented by a 'Flag'

[34] Ofcom, Safety measures on video-sharing platforms survey (quantitative research) (2021)
[35] Ofcom, Online Nation: 2022.
[36] Ofcom, Safety measures on video-sharing platforms survey (quantitative research) (2021)
[37] An ellipsis is presented on platforms as '…'.

icon.[38] The reporting process can also differ widely between platforms. Permutations include: categories of harm that are suggested to users; how the harms are described; the order in which those categories are presented; and the number of the categories of harm. Some VSPs also allow users to report and flag non-video content (e.g., comments and direct messages).

3.47 Given the range of behavioural factors that could be relevant to assessing the effectiveness of reporting mechanisms, we choose to focus primarily on barriers that might prevent users from starting the reporting process in the first place. Based on the principles of the COM-B 'Behaviour Change Model'.[39] Table 1 below sets out examples of the kinds of behavioural barriers which may prevent people from engaging with online reporting mechanisms.

Table 1: Examples of behavioral barriers to effective online reporting mechanisms

Barrier	How it relates to reporting mechanisms
Attention	Users may be too distracted or preoccupied by the ongoing activity to report content.
Awareness	Although users were generally aware of reporting mechanisms, they may be unaware of what they can/should be reporting.
Resources & time	Users may perceive the reporting process to take too long and decide not to start a report.
Prompts in the environment	The reporting mechanism may not be prominent enough to prompt user action.
Role models	There may not be active role models who promote user reporting.
Norms	Although the action of reporting is private, it may not be a norm to report potentially harmful content. This may be particularly true amongst younger users.
Belief of consequences	Users may be unsure about the consequences of reporting, specifically whether reporting is likely to have an impact.
Accountability	Users may not feel responsible for reporting content and are not held account for their inaction.
Identity	Reporting content may not fit with their (online or offline) identity.

[38] Ofcom, 2021, Video sharing platform guidance.
[39] Michie, S., van Stralen, M.M., and West R. (2011) The behaviour change wheel: a new method for characterising and designing behaviour change interventions.

Emotions	Users may only decide to report when they are personally emotionally affected.

3.48 Table 1 demonstrates that there could be a wide range of behavioural factors which are constraining users' ability to report content that they are concerned about online. However, not all are capable of being tested in our experimental set-up. For instance, issues around identity, role models and accountability are more complex issues to explore in one-off experimental research.

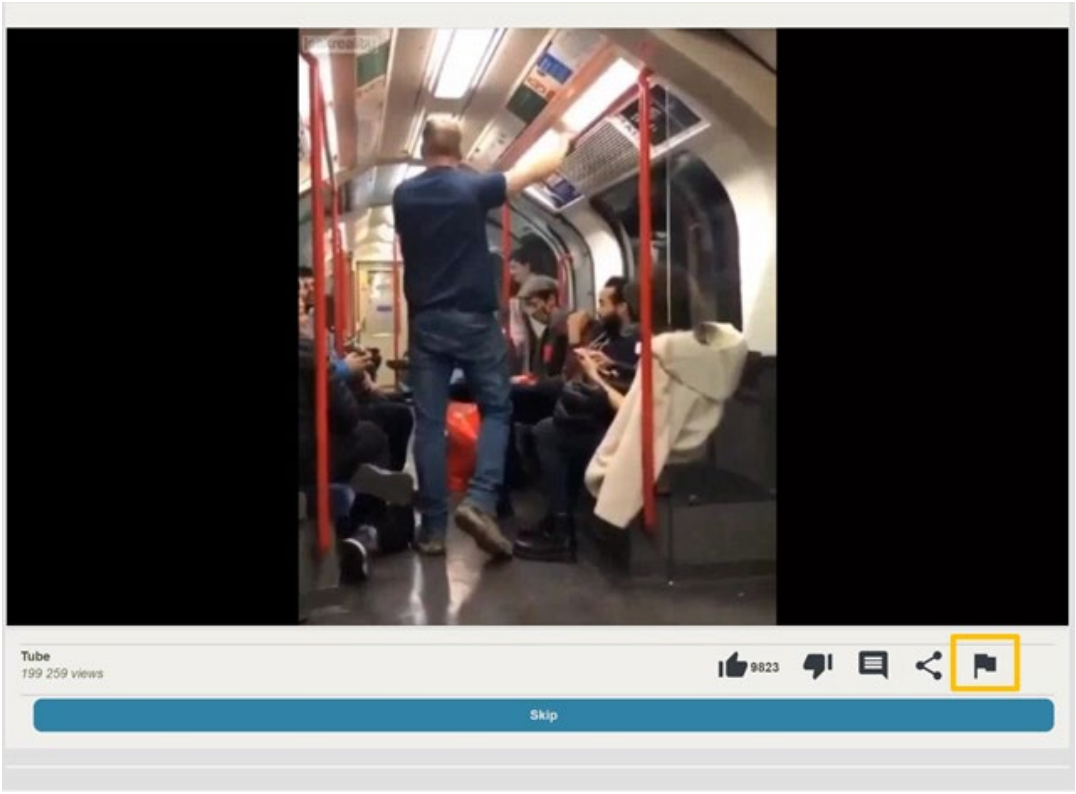
Interventions we tested

3.49 As with the alert messages research, to select the interventions for our research, we used our review of the behavioural literature and observational research and chose interventions which (a) are already used by some VSPs, (b) had limited research in the public domain on their effectiveness at improving safety, and (c) could be replicated in our experimental setting. Using these criteria, our first two interventions involved increasing the salience of the reporting mechanism. In the third intervention, we retained the features of the second intervention and simplified the reporting process to reduce the effort involved in submitting a report.

3.50 As with the alert messages research, the main interface in the control was as in Figure 1 (in the 'Experimental design' section above).

3.51 **Salience** – Because people have limited attention and (potentially) awareness, the salience of options can serve to raise awareness of the reporting function. Increased salience can also prompt people to take action. To increase salience of the reporting mechanism, we moved the reporting option from the menu behind the ellipsis and inserted a 'flag' icon on the main options bar – see Figure 5 below.

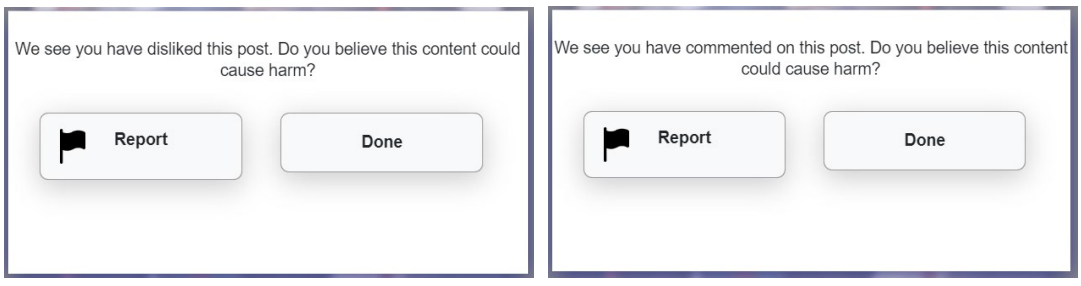
Figure 5: Increasing the salience of the reporting mechanism



Note: The yellow box highlights the change in the user interface relative to the control (in the control group the flag button was behind an ellipsis). The yellow box did not appear in the user interface as viewed by participants.

3.52 **Salience plus prompt** – the behavioural barriers listed in Table 1 suggested that users may feel that content is inappropriate but remain unsure as to whether they should report it and so potentially signal their concerns in other ways. To encourage action in this group, we included an additional prompt *when a research participant commented or disliked the content*. The prompt asked participants if they would like to report the content – see Figure 6 below.

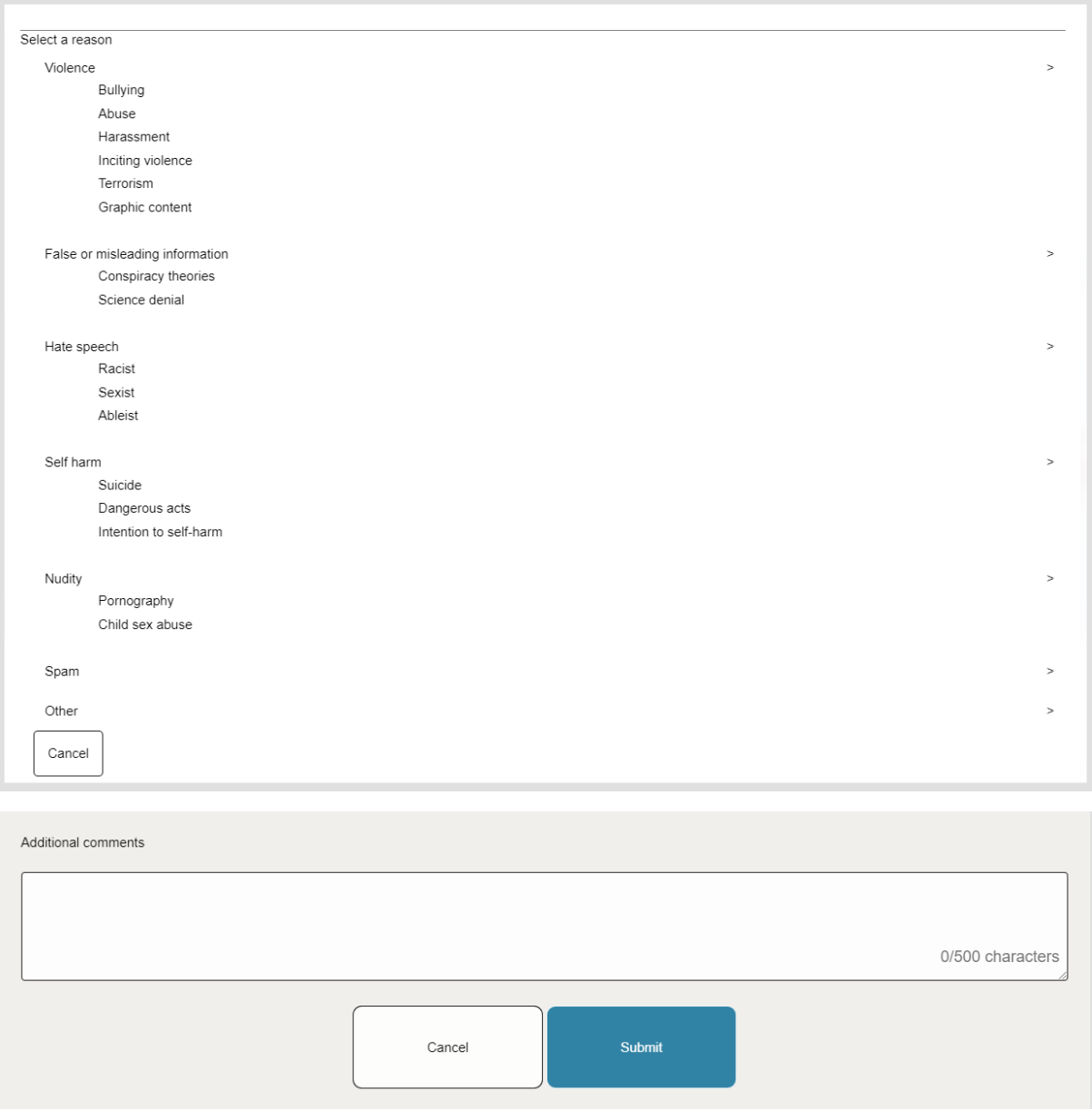
Figure 6: 'Salience plus prompt' reporting mechanism



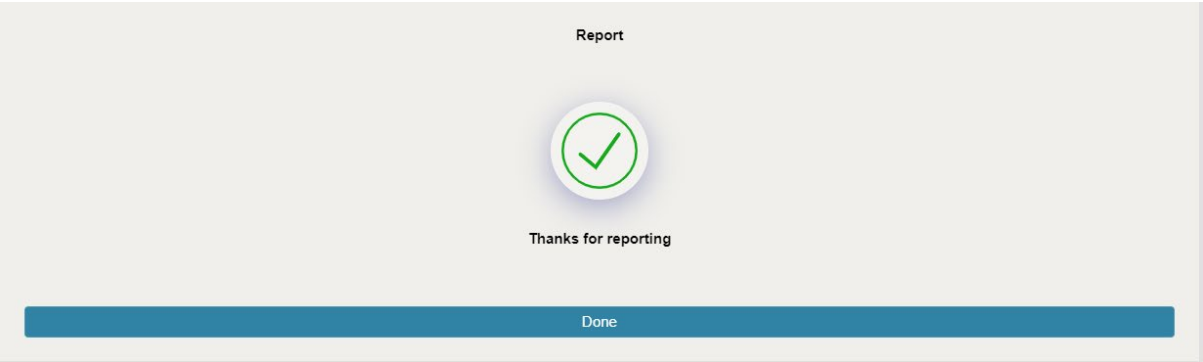
Note: The pop-up on the left was shown to people who 'disliked' content. The pop-up on the right was shown to people who commented on content.

3.53 **Salience plus prompt plus simplification** – People are thought to be more likely to engage in a behaviour if it is easy to do.[40] In addition to the previous interventions, our final intervention redesigned the reporting process to make it simpler. The aim was to shorten the time taken to report, as well as reducing the cognitive effort required.

Figure 7: Simplified reporting mechanism



[40] Behavioural Insights Team (2014), EAST: Four simple ways to apply behavioural insights.



Primary and secondary outcomes

3.54 The **primary outcome** measure from this research was the number of reports submitted when participants viewed legal but potentially harmful videos.

3.55 The **secondary outcomes** from this research were:

a) The number of reports submitted for 'neutral' content i.e., content that should not have been flagged. We included this measure to understand if raising the prominence of reporting mechanism or simplifying the reporting process led participants to 'over-report'. That is, to report content that is highly unlikely to be considered harmful.

b) The number of reports of harmful content started but not finished and the number of reports of neutral content started but not finished. We included this measure to understand if the reporting process was challenging for participants to complete, even if the interventions we tested caused more participants to start reporting.

c) The number of times participants skipped harmful content and the number of times participants skipped neutral content.

d) The number of reports submitted for harmful content that accurately categorised the harmful content according to the options available (i.e., a video containing violent content categorised as contained violent content).

3.56 We also asked research participants to complete an additional task at the end of the experiment. Research participants were asked how they would categorise an additional video clip of legal but potentially harmful content, even if they would not have reported it. The focus of this exercise was to test the accuracy of reporting according to the treatment arm.

Ethical considerations

3.57 As part of our research design, we carefully considered ethical concerns about exposing research participants to legal but potentially harmful content. We worked with Kantar

Public to establish a process to ensure the wellbeing of participants in the research. Key features of this approach were:

a) The video content used in the experiments did not include illegal content. The experiment tested the effectiveness of the chosen safety measures in relation to content that Kantar considered to be 'legal but potentially harmful content'. That is, content that was legal but that, in Kantar Public's assessment, some people might consider offensive or problematic.

b) There was a review of the video content by an internal committee within Kantar Public to check that the content that had been submitted was not illegal and that Kantar Public was comfortable about exposing participants from its Lifepoints panel to the content.

c) Kantar Public has both corporate and individual membership of the industry body - the Market Research Society ('MRS') - and the research it carries out is therefore governed by the MRS professional MRS Code of Conduct. In relation to ethical research, the Code of Conduct requires that "Members must take all reasonable precautions to ensure that participants are not harmed or adversely affected by their professional activities and ensure that there are measures in place to guard against potential harm".[41]

d) Kantar Public also subscribes to the ESOMAR code of conduct which includes the general principle that researchers will behave ethically[42] and not do anything which will harm a data subject as well as a specific 'Duty of Care' responsibility which includes: "Researchers must exercise special care when the nature of the research is sensitive or the circumstances under which the data was collected might cause a data subject to become upset or disturbed".[43]

e) Participants in the trial had to be 18 years of age or over. Anyone under 18 was screened out at the selection stage.

f) At the screening stage, potential research participants were also informed that some of the videos that they would see might show violence, extreme views, or harmful content and they had the option not to participate in the trial if they did not want to be exposed to that content.

g) During the trial itself research participants were not required to watch any of the videos and were free to skip any video at any time and were not penalised for doing so. In addition, research participants could leave the trial at any point.

h) Research participants were provided with a debrief screen at the end of the experiment. The screen provided web links to support related to the content shown in the experiment.

---

[41] No.9 General Rules of Profession Conduct, MRS Code of Conduct 2019.
[42] ESOMAR Code on Market and Social Research
[43] Section 9.3 ESOMAR Guideline on duty of care

3.58 Kantar Public's Profiles' Privacy team ensured that the research process complied with the relevant regulations such as the UK GDPR, and industry best practice. Kantar Public also adhered to the Market Research Code of Conduct (2019). We also reviewed the approach and considered it appropriate.

4. Results and insights

4.1 In this section we set out the high-level findings of the two experiments, discuss our insights from the results, and the limitations of our research. We also set out ideas for further research based on these findings. A more detailed description of the findings and the supporting statistical analyses is set out in Appendices 1 and 2. Whilst we present charts below to demonstrate our results, the appendices contain the more detailed results of the underlying analyses, showing the detail of statistical analysis that was conducted. Unless otherwise stated, by convention our test for statistical significance is at the 5% level (i.e., $p<0.05$).

Alert messages

4.2 In the course of more recent work, it was established that there was an inconsistency in the data collection of skipping behaviour in certain circumstances. That is, if, after choosing to watch a legal but potentially harmful video, participants engaged with a video in any way (i.e. like, dislike, comment, share, report) and then decided to skip to the next video before the completion of the current video, they were recorded as having watched the entire video, when in subsequent Ofcom trials they were recorded as having skipped the video.

4.3 To make our results consistent across trials, we worked with Kantar Public to identify an alternative method of classifying whether these participants had skipped or not based on the view time of the video.[44] The following results are based on this alternative method of classification.

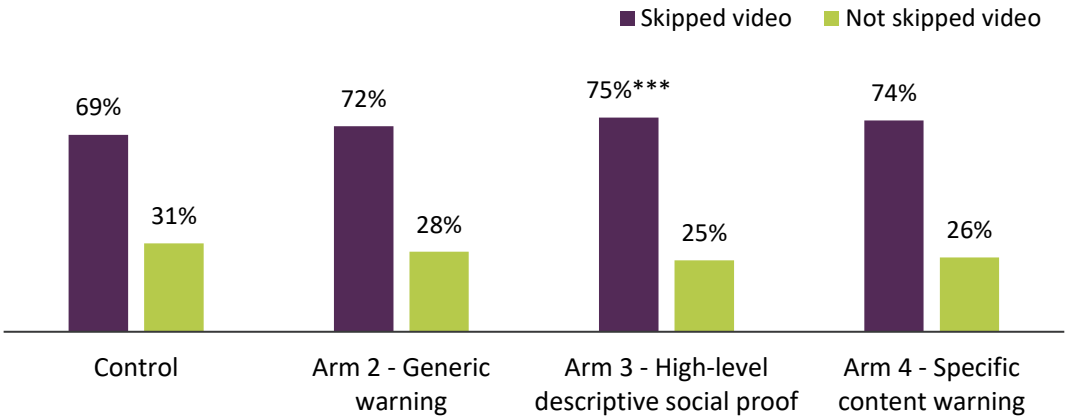
Results

Primary outcome: only the alert with the high-level descriptive social proof message increased the likelihood of skipping legal but potentially harmful videos

4.4 We found that only the alert message containing a high-level descriptive social proof message led to a statistically significant increase in the overall level of skipping legal but potentially harmful videos compared to the control arm. In our experiment the proportion of legal but potentially harmful videos skipped was 69% in the control arm, 72% in the generic warning arm, 75% in the high-level descriptive social proof arm, and 74% in the specific content warning arm, as shown in Figure 8.

[44] More detail can be found in section 5 of the accompanying trial report.

Figure 8: Proportion of legal but potentially harmful videos skipped by participants

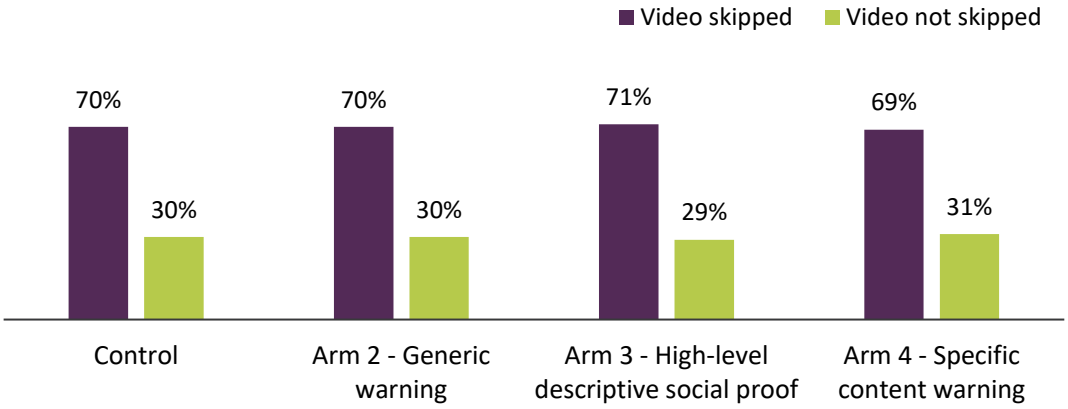


*Note: *** denotes statistical significance at the 0.1% level (p<0.001) compared to the control arm.*

Secondary outcome (1): alert messages did not increase the likelihood of skipping neutral content

4.5 We found that an exposure to alert messages for legal but potentially harmful videos did not led to an increase in the likelihood of skipping neutral videos in the treatment arms compared to the control arm. This effect is illustrated in Figure 9 below which shows no significant differences in the percentage of neutral videos skipped by participants across the three treatment arms.

Figure 9: Percentage of neutral videos skipped by research participants



Secondary outcome (2): participants choosing to watch legal but potentially harmful videos watched for longer

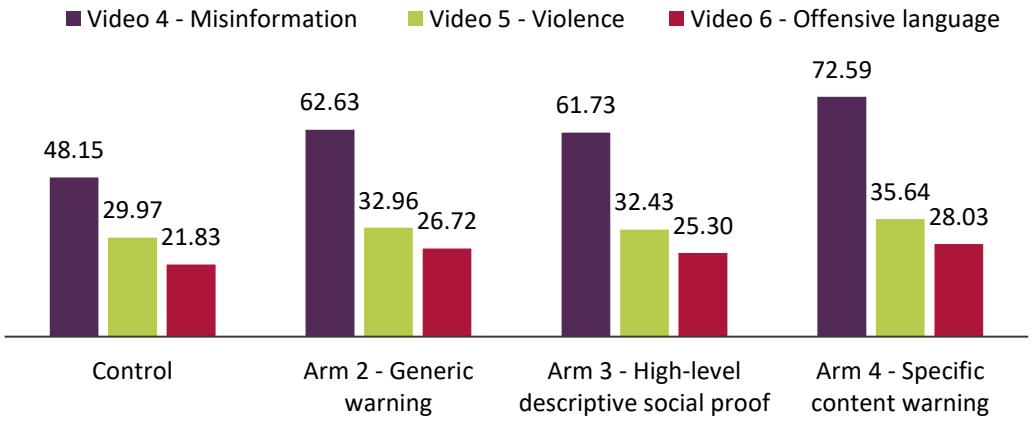
4.6 Participants in the treatment arms who chose to watch the legal but potentially harmful videos watched the videos on average for significantly longer than those in the control arm.[45] That is, having been exposed to an alert message, if participants chose to watch a

[45] The results for the three treatment arms were all significant at the <0.01 significance level.

legal but harmful video, they watched that video for longer on average than participants in the control arm.

4.7 Figure 10 below shows the increase in viewing times for all three legal but potentially harmful videos across the three treatment arms, when compared with the control arm.

Figure 10: Average viewing times (in seconds) for research participants who chose to watch the legal but potentially harmful videos



4.8 Figure 10 shows that the biggest increase in average viewing time across all three videos was in the specific content warning arm when compared with the control arm. For instance, the average viewing time for Video 4 (misinformation) increased from around 48 seconds in the control arm to around 73 seconds in the specific content warning treatment arm.

Secondary outcome (3): there was a reduction in the reporting of legal but potentially harmful content

4.9 We also assessed whether the use of alert messages might lead to an increase in the reporting of legal but potentially harmful content. For instance, having been warned about the content they were about to see, research participants might then be 'primed' to be more inclined to report it. In fact, the probability of reporting the legal but harmful videos was statistically significantly lower in the high level descriptive social proof warning arm and specific content warning arm than in the control.

Post-survey attitudinal questions

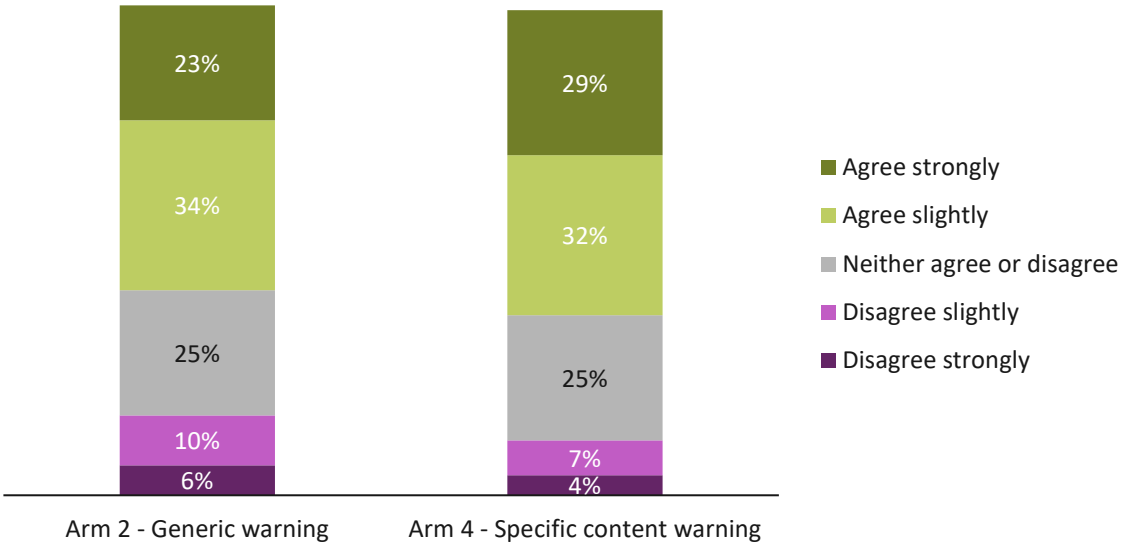
4.10 As part of the questions which followed the main trial, we also asked research participants that were exposed to the alert messages how useful and how annoying they found the alert message that they were exposed to and, if they had watched a video, whether they regretted it.

A majority of participants found alert messages to be useful

4.11 The majority of respondents for each trial arm agreed that they found the alert messages to be useful.[46] In contrast, only a relatively a small percentage of respondents reported that they did not find the alert messages to be useful.[47]

4.12 In particular, research participants who saw the specific content warning message were significantly more likely to report that they found the alert message useful compared to those participants who saw the generic warning message. This is illustrated in Figure 11 below. The figure shows that 57% of participants in the generic warning arm slightly agreed or strongly agreed that the alert messages were useful to them. A higher share of participants (61%) who saw the specific content warning slightly agreed or strongly agreed that the alert messages were useful to them.

Figure 11: Participants attitudes towards the usefulness of alert messages in the generic warning and specific content warning arms[48]



4.13 There was no significant difference in usefulness between the generic warning and high-level descriptive social proof warning arms.

[46] Between 57-62% of respondents (depending on the trial arm) agreed or strongly agreed with the statement that 'I found the warning messages I just saw useful to me.'

[47] Between 11-16% of respondents (depending on the trial arm) disagreed or strongly disagreed with the statement that 'I found the warning message I just saw useful to me'.

[48] "To what extent do you agree or disagree with the following statement: 'I found the warning messages I just saw useful to me.'" Does not sum to 100% due to rounding.

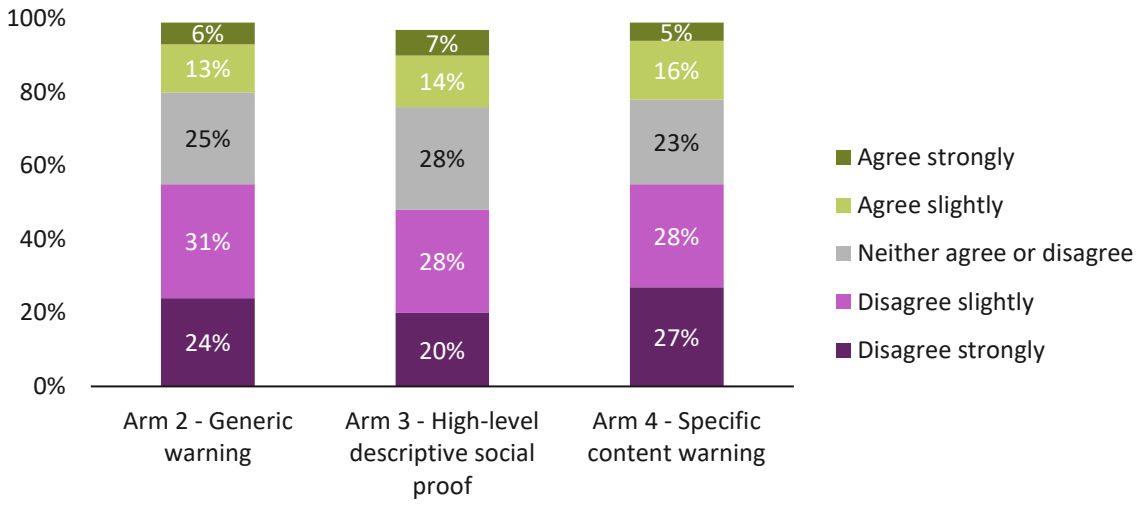
Few participants found the alert messages to be annoying

4.14 We asked whether participants who had been exposed to alert messages found them annoying. If participants found the warning messages to be annoying there could be the risk that any impact from the warning messages would be quickly eroded over time.

4.15 As shown in Figure 12, in general participants who were exposed to the alert messages did not report that they found them to be annoying.[49] Figure 12 shows that across all three treatment arms, only around 20% of research participants agreed with the statement that they found the alerts to be annoying while around 50% of research participants disagreed with that statement. 23-28% of research participants did not express a view.

4.16 This does not mean that the impact of the warning messages would not decline over time – users may simply get used to them - but it does tend to suggest that the alert messages we tested did not irritate users.

Figure 12: Participants attitudes towards the 'annoyingness' of alert messages across treatment arms



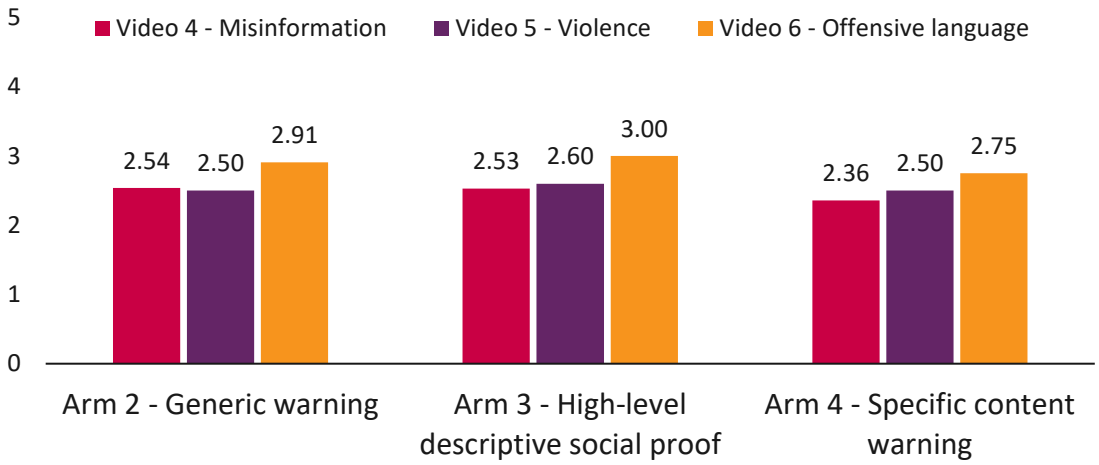
Participants did not regret choosing to watch a video after receiving an alert message

4.17 Finally, we asked whether research participants who saw alert messages but chose to watch the videos regretted their decision.

4.18 We found that there were no significant differences between the level of regret expressed by participants between the different treatment arms over the legal but potentially harmful videos. This is illustrated in Figure 13 below.

[49] Between 19-21% of respondents (depending on the trial arm) agreed or strongly agreed with the statement that 'I found the warning messages I have just seen annoying.'

Figure 13: Average levels of regret at having watched different videos[50]

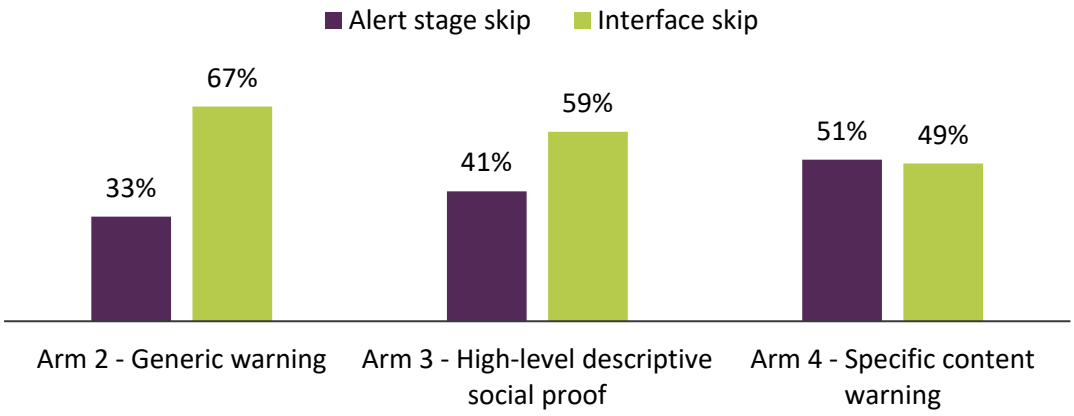


Alert messages had different impacts on when participants skipped

4.19 Our primary analysis focuses on overall levels of skipping – that is if participants skipped a video at either the alert stage, or the interface stage i.e. once a video had started to play. Although we must be cautious with the interpretation of the following analysis, it appeared that the alert messages had different impacts on *when* participants chose to skip.

4.20 Figure 14 below shows that the specific content warning led to the highest level of skipping at the alert stage, compared to the other treatment arms.

Figure 14: Proportion of skips occurring at the alert stage or the interface



Discussion of results: alert messages

4.21 **Our research suggests that alert messages which include a high-level descriptive social proof message can be effective at enabling users to make more informed viewing**

[50] "To what extent do you agree with the following statement for that video: 'I regretted watching the video.'?" on a scale of 1 (Disagree strongly) to 5 (Agree strongly).

choices. We found that exposing research participants to an alert message which included a high-level descriptive social proof message before videos containing legal but potentially harmful content increased the likelihood that those videos would be skipped by research participants, relative to the control arm.

a) We had expected the specific content warning alert message to have a significant impact on behaviour compared to the control arm, as well as compared to the generic warning, because it gave research participants more detailed information about the nature of the video content they were about to see and so enabled a more informed decision. However there was no statistically significant difference between the relevant arms.

b) Further analysis indicates that alert messages had different impacts on the point at which research participants skipped the legal but potentially harmful videos. In the specific content warning treatment arm a significantly higher proportion of the legal but potentially harmful videos were skipped at the alert message stage compared to the other two treatment arms.

c) We did not find any evidence to suggest that the alert messages led to an increase in the skipping of neutral videos.

4.22 **Participants who were exposed to alert messages about upcoming videos but who chose to watch the video content were more likely to be comfortable with that content.** Participants in our research who were exposed to alert messages, but then continued to watch the legal but potentially harmful videos, were not only less likely to report the legal but potentially harmful content but also were more likely to watch the videos for longer on average.

a) This pattern of behaviour suggests that the alert messages were effective at giving participants who might have been uncomfortable with the content in a video the information and prompt to skip that video. Those participants who went on to watch a video having been warned about the content would also be better placed to understand the risk of harm: they were either comfortable with the content or were alive to the fact that they might find the content problematic and were 'primed' to skip away from it if necessary.

b) However, this pattern of behaviour could also be because research participants did not want to believe that they have made a bad decision and developed a psychological defence mechanism to offset the cognitive dissonance. This could lead them to watch the legal but potentially harmful videos for longer than they might otherwise have chosen to in the absence of an alert message.

c) On balance, we consider that the former explanation is the more likely to be the case. This is because:

    i) The participants in our research reported that they found the alert messages to be useful, and that they did not find them to be annoying.

ii) We did not find any significant differences between the level of regret at having watched the different videos across the different treatment arms. This tends to point towards the alert messages having been effective at prompting participants to make an informed viewing choice either at the alert stage or once a video had started to play.

4.23 **Alert messages that contain specific information about the type of legal but potentially harmful content are more useful and less annoying to viewers.**

a) Although we only found a statistically significant impact on the likelihood of skipping legal but potentially harmful content for the alert with the high-level descriptive social proof message, research participants reported that they found the specific content warning alert message to be both more useful and less annoying than the other forms of alert messaging.

b) We found that the specific content warning message led to more participants skipping legal but potentially harmful videos at the alert message stage i.e., without participants watching them at all. This could suggest that specific warnings are more effective at helping users avoid legal but potentially harmful content altogether compared to the other forms of alert messaging.

Potential areas for further research

4.24 **Testing the impact of combining alert messages with a default action for the users to skip the videos.** Our alert messages did not have a strong default action for users to take. That is, research participants were presented with a choice between a 'skip video' button, and a 'view video' button but neither one was pre-selected. In some cases where platforms are currently making use of alert messages, the message is combined with a default option of watching the video (see Annex 1). Given the importance of default settings, it is possible that in this case having the default option to 'view video' could offset any impact from the alert message.

4.25 There is potentially value in exploring the impact of combining different alert messaging with a default option set to "skip" the video rather than to "view" the video.

4.26 **Testing other forms of alert messages.** The alert messages we tested in our experiment were based on the types of alert messages currently used by platforms. One way to develop this research would be to consider other, potentially stronger forms of messaging. For example, messages which put more emphasis on skipping content leading to a reduction in the overall circulation of the content or on drawing attention to over-confidence on the part of the user about their ability to avoid negative aspects of the content. We could also look to test stronger forms of social proof messages (e.g., "People you follow found this content to be sensitive.").

4.27 We could also explore whether **introducing a short delay between the alert message and the video starting** to play has an impact on users' skipping legal but potentially harmful

content. However, we are aware that other research suggests that users find delays to be annoying and – in the case of research for TikTok – may not have a significant impact on viewing (see Annex 1).

Reporting mechanisms

Results

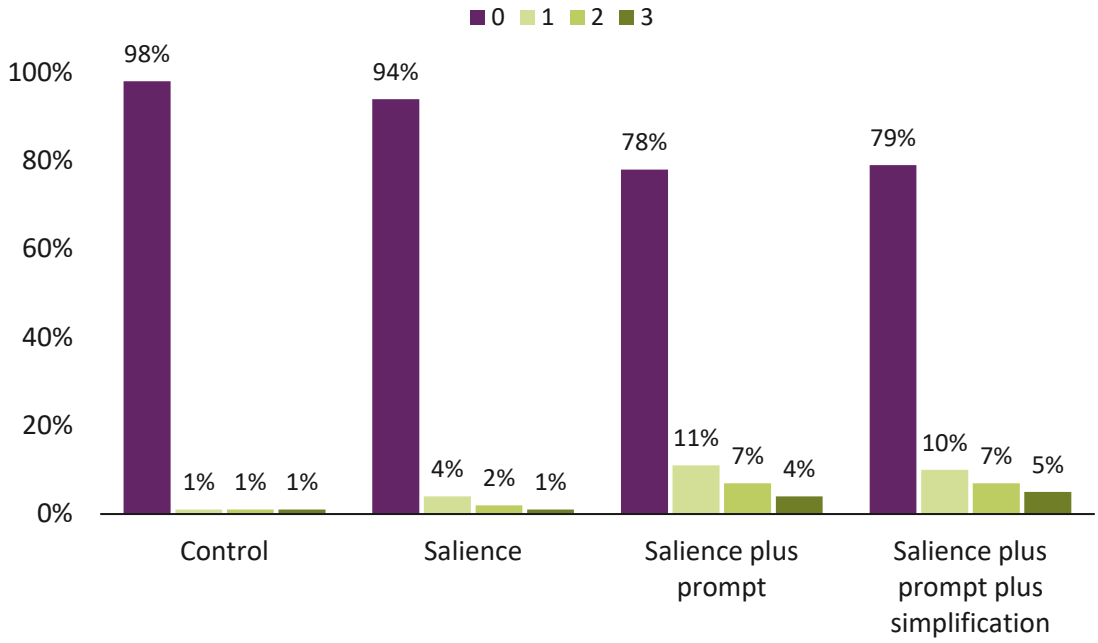
Primary Outcome: raising the prominence of the reporting function increased the likelihood of reporting legal but potentially harmful content

4.28 We found that all three of the interventions to encourage people to report legal but potentially harmful content performed significantly better than the control arm. [51,52] Increasing the prominence of the reporting function plus a prompt and increasing the prominence of the reporting function plus a prompt plus simplifying the reporting process also had a statistically significant impact on the volume of reporting compared to just raising the prominence of the reporting mechanism on its own.[53]

4.29 This is illustrated in Figure 15 below. The figure shows that levels of reporting of the legal but potentially harmful videos in the control arm were very low with 98% of research participants choosing not to report any legal but potentially harmful videos. In contrast, in the treatment arms, the percentage of research participants who did not report any of the legal but harmful videos fell (the purple bars in the treatment arms decline relative to the purple bar in the control arm) and there was an increase in the percentage of participants reporting the legal but potentially harmful videos (the green bars in the treatment arms are higher than in the control arm). By way of example, 1% of participants reported one legal but potentially harmful video in the control arm, compared to 10% of participants in the salience plus prompt plus simplification treatment arm.

[51] See Table 2 in the accompanying trial report. All three treatment arms are significantly different from the control arm at the 0.1% significance level ($p<0.001$).
[52] We also note that the direction of these effects was not sensitive to the inclusion of age and the order in which videos were seen by research participants as covariates.
[53] See Table 3 in accompanying trial report. There was a statistically significant difference between the salience plus prompt and salience plus prompt plus simplification treatments arms compared to the salience treatment arm at the 0.1% significance level ($p<0.001$).

Figure 15: Percentage of participants reporting legal but potentially harmful content



4.30 We did not find any statistically significant difference in the overall probability of reporting legal but harmful videos between the salience plus prompt and the salience plus prompt plus simplification treatment arms.

Secondary outcome (1): we did not find evidence of an increase in over-reporting

4.31 As part of the secondary analysis, we considered whether making the reporting function more salient or prominent led to over-reporting of neutral content. However, the actual number of over-reports of the neutral content in each of the treatment arms was very low.[54] As a result, it was not possible to carry out reliable statistical analysis on these results and we cannot draw any statistical inferences in relation to the apparent increase in over-reporting in the salience plus prompt plus simplification treatment arm can be drawn.

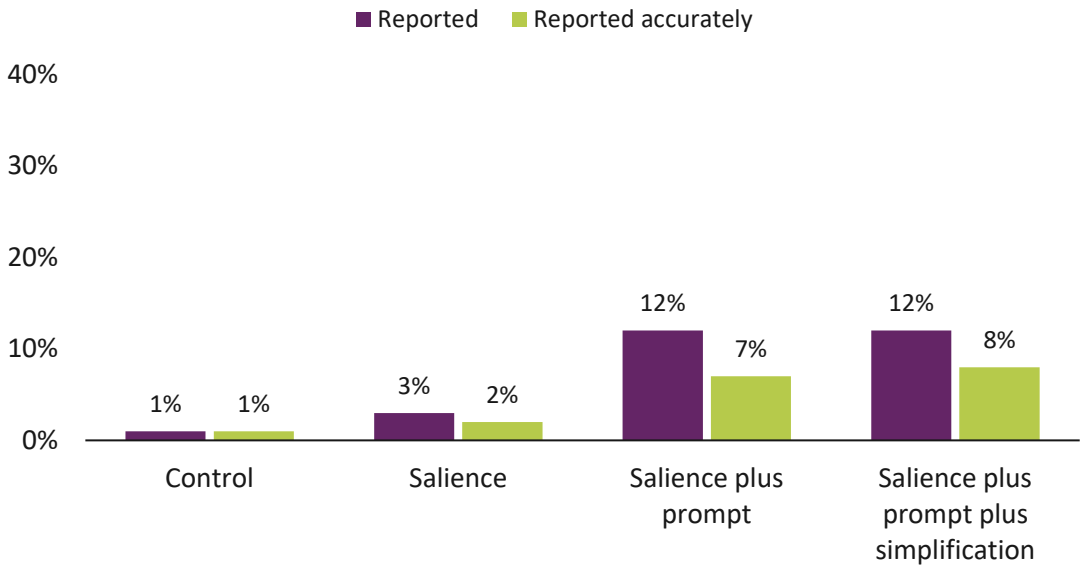
Secondary outcome (2): we did not find evidence of a reduction in the accuracy of reporting

4.32 We found that, of all the videos that could have been reported, in the salience plus prompt and the salience plus prompt plus simplification interventions arms, there was a statistically significant increase in the likelihood of legal but potentially harmful videos being accurately reported compared to the control arm. [55] That is, research participants not only reported more legal but potentially harmful videos but they also categorised the videos correctly. Figure 16 below shows that in the salience plus prompt and the salience plus prompt plus simplification treatment arms, 12% of all the videos that could have been

[54] There were just two over-reports in each of the control, salience, and salience plus prompt arms. There were 12 in the salience plus prompt plus simplification arm.
[55] Statistically significant at the 0.1% significance level ($p<0.001$).

reported as legal but potentially harmful content were reported compared to just 1% in the control arm.[56] In addition, the number of accurately reported videos increased from 1% of all the videos that could have been reported in the control arm to 7-8% of videos in the salience plus prompt and salience plus prompt plus simplification arms.

Figure 16: Percentage of videos accurately reported and categorised as legal but potentially harmful content.



4.33 Both the salience plus prompt and the salience plus prompt plus simplification treatment arms also resulted in a lower likelihood of inaccurately reporting legal but harmful videos compared to the control and the salience treatment arms. There was, however, no statistically significant difference in the odds of inaccurate reporting between the salience plus prompt and salience plus prompt plus simplification treatment arms.[57]

Secondary outcome (3): we did not observe any change in skipping behaviour

4.34 We also assessed whether any of the treatments prompted a change in research participants' propensity to skip videos.

4.35 Skipping videos was already a common behaviour in the control arm, with only around 25-30% of research participants choosing not to skip the legal but potentially harmful videos.

4.36 We did not find that any of the treatments had a statistically significance impact on research participants' propensity to skip the legal but potentially harmful content relative

[56] In each treatment arm there could be up to 1800 reports of legal but potentially harmful videos i.e., 3 legal but potentially harmful videos x 600 research participants.
[57] Additional analysis by Ofcom indicates that the accuracy rate (i.e., the number of accurate reports of legal but potentially harmful content as a proportion of total reports) was significantly higher in the treatment arm which included the simplification of the reporting process compared to the salience plus prompt treatment arm.

to the control arm. That is, increasing the prominence of the reporting function did not lead to participants choosing to skip the legal but potentially harmful content.

4.37 We also did not find any statistically significant difference between the propensity of research participants to skip the neutral content videos compared to the legal but potentially harmful videos across the different treatment arms.

Secondary outcome (4): a lack of incomplete reports suggests navigating the reporting process was not an issue

4.38 There was not enough data to conduct an analysis of the differences in the number of incomplete reports between the different treatment arms because there were no incomplete reports in the control, salience and salience plus prompt arms. In the salience plus prompt plus simplification arm, there were five incomplete reports. The low number of incomplete reports tends to suggest that navigating the reporting process was not a problem for research participants.

Post-trial reporting accuracy

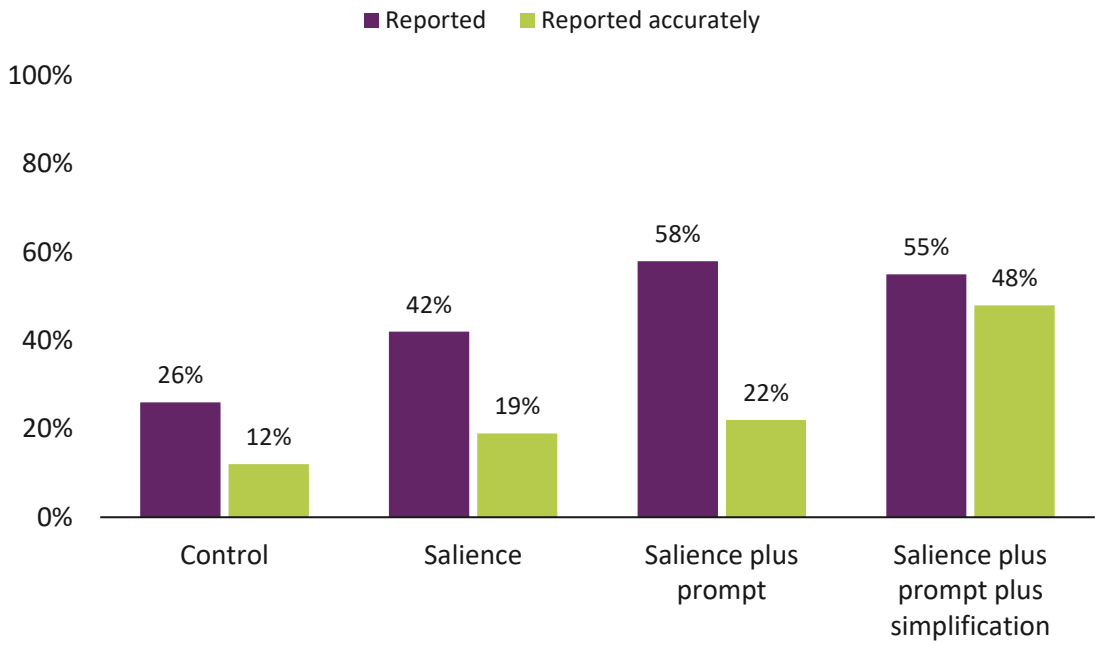
4.39 As set out in section 3, at the post-trial stage we also asked research participants to carry out an exercise which set out a scenario in which they were asked to submit a report of a legal but potentially harmful video. This allowed us to explore further the accuracy of reporting by trial arm.

4.40 In this exercise there was a statistically significant increase in the likelihood that participants categorising the video correctly in all three treatment arms compared to the control arm.[58] Specifically, all the treatments increased the odds of accurately reporting the legal but potentially harmful content compared to the control arm.

4.41 Figure 17 below shows the percentage of participants across each trial arm that reported the follow-up video and the percentage that reported it correctly. In this scenario the salience plus prompt plus simplification treatment arm led to a statistically significant increase in the percentage of participants that correctly categorised the content compared to control and the other two treatment arms.[59]

[58] Statistically significant at the 0.1% significance level ($p<0.001$).
[59] Statistically significant at the 0.1% significance level ($p<0.001$).

Figure 17: Percentage of participants reporting the follow-up video and accurately reporting the content of the video.



Discussion of results: reporting mechanisms

4.42 **Making the reporting function more prominent led to an increase in the likelihood of reporting of legal but potentially harmful content**

a) In the control arm, the level of reporting of legal but potentially harmful content was very low: only around 2% of research participants reported any of the legal but potentially harmful videos. Ofcom's consumer survey research has indicated higher levels of reporting of potentially harmful by internet users (see Section 3).

b) The low levels of reporting in the control arm could be in part due to choices about the content that was categorised as legal but potentially harmful and used in the experiment. For instance, for ethical reasons, the content in our research was intended to less contentious and provocative and to avoid certain topics (e.g., self-harm, bullying, harassment) that could be regarded as particularly emotive.[60] At the same time, the higher levels of reporting in Ofcom's consumer research could be a result of participants being asked about reporting content across *all* their browsing experiences rather than in relation to specific types of content.

[60] The low level of reporting was unlikely to be related to research participants being passive and simply not engaging with the content. We found that research participants engaged with both the neutral and the legal but potentially harmful content by liking / disliking, commenting etc. For instance, around 23-27% research participants disliked the legal but potentially harmful videos (whereas 9-13% of participants liked the same content).

c) Against this backdrop, simply promoting the salience of the reporting function (from being behind an ellipsis ('…') to a flag icon on the main options bar) prompted a statistically significant increase in the overall likelihood of reporting legal but potentially harmful content. We saw the percentage of participants reporting one legal but potentially harmful video increasing from 1% in the control arm to 4% in the salience intervention – a fourfold increase. In addition, the percentage of participants reporting any legal but harmful video increased from 2-3% of participants in the control arm to 7% of participants in this intervention.

d) Our other interventions both introduced an additional variation to the user interface i.e., a targeted active prompt when the participant disliked or commented on a video and a simplification of the reporting process.

e) Similar to the alert messages in our first experiment, the targeted prompt required research participants to make an active choice to report or not.[61] Raising the profile of the reporting function combined with an additional prompt to report for participants who disliked or commented on legal but potentially harmful content had a more significant effect on increasing the likelihood that the content would be reported. For instance, with this intervention 11% of research participants reported one of the legal but potentially harmful videos compared to 1% in the control group. Furthermore, 4% of research participants reported all three legal but potentially harmful videos compared to 1% in the control group.

4.43 **There was no significant difference in the likelihood of reporting legal but potentially harmful content between our salience plus prompt intervention and our intervention which included simplifying the reporting process.**

a) We had expected that simplifying the reporting process might increase the likelihood of participants reporting legal but potentially harmful content by reducing the effort involved in making a report. Although research participants would need to make at least one report in order to experience the simplification in the reporting process, we then expected to see an effect in terms of an increased propensity to report content the next time a participant came across content they found offensive. At this stage, we have not been able to conduct a more detailed analysis to establish whether any effect could be detected.

b) As indicated above, the very low levels of incomplete reports do not suggest that participants found the reporting process itself to be complex or difficult to navigate.

c) From the secondary analysis, there is some evidence to suggest that simplifying the reporting process could help to increase the accuracy of reporting. This is discussed in more detail below.

[61] The use of this active prompt mirrored the alert messages we tested in alert messaging trial. That is, a prompt required the research participant to make an active choice about reporting and the alert messages we tested also required the participant to make a choice between viewing or skipping the video.

4.44 **Research participants appeared to use proxies - such as the "dislike" and "comment" functions - to register their concern about content rather than using the reporting function.** It is possible that participants found it easier / more convenient to use the functionality immediately available to them (the dislike or comment buttons) to register concerns about content rather than going to the effort of searching for the reporting function. It is also possible that reporting content is perceived as being a more formal step or process or that users don't naturally self-identify as someone who reports content.

4.45 Our research suggests that participants using other functionality as a proxy for reporting can be encouraged to use the reporting process by including appropriate prompts. Once the user starts the reporting process, they do not appear to struggle to complete a report about content they are concerned about.

4.46 **Increasing the prominence of the reporting function does not necessarily lead to a reduction in the accuracy of reporting of legal but potentially harmful content.**

a) In informal discussions, platforms have expressed the concern that making it easier to report content will lead to an increase in spurious or even vexatious reports and that dealing with such reports would be an inefficient use of their resources and detract from genuine complaints.

b) Our research suggests that in fact the volume and accuracy of reporting of legal but potentially harmful content can increase when the reporting function is made more prominent. There is some evidence to suggest that a combination of making the reporting mechanism more prominent combined with a well-designed reporting process can lead to an increase in both the volume and accuracy of reporting of legal but potentially harmful content.

Potential areas for further research

4.47 There are a number of aspects of our experimental set-up where we believe there could be scope for further testing. For instance, in our research we used video clips that were typically in the order of 30-45seconds. It is possible that users may behave differently when exposed to longer or shorter videos depending on the circumstances. For instance, in the case of shorter video clips, there would be less time to consider reporting content before the next video started to play.

4.48 We are also aware that the experiment was set up to play six videos – comprising three neutral videos and three videos of legal but potentially harmful content – in a random order. This was obviously an approximation of a user's typical viewing pattern in terms of the prevalence of legal but potentially harmful material for the purposes of our experiment. As a result of this experimental set-up, some participants could have been exposed to three legal but potentially harmful video clips in a row and this could have had a bearing on the likelihood of their reporting this content. It is not possible to anticipate in advance whether this could have made them more or less likely to report legal but

potentially harmful content. However, from an inspection of the experimental results, it does not appear that such participants reacted differently to other participants.

4.49 We consider that could also be interesting to investigate the impact of changing the ratio of neutral to legal but potentially harmful content on the probability of reporting content but this would need to be informed by statistics on the prevalence of users coming across legal but potentially harmful content in a typical browsing scenario.

4.50 Finally, our research indicates that research participants did not find it difficult to navigate the reporting process – the number of incomplete reports was extremely low across all trial arms even with the increase in reporting volumes. Our research does suggest that the accuracy with which legal but potentially harmful content is categorised does not have to suffer from an increase in reporting. We consider that it would be interesting to explore this issue in more detail and to assess how the choice architecture around the reporting process could be changed to support the accurate categorisation of legal but potentially harmful content.

Limitations of online RCTs

4.51 There are a number of limitations to the online RCTs we carried out. The main limitations to our research approach are:

a) We observed users' behaviour in a research environment i.e., users were recruited by a research agency and were interacting with a mock-up of a VSP. The mock-up of the VSP interface did include the ability to play actual video content and research participants could interact with the functionality found on many VSPs e.g., likes / dislikes, share, comment, report etc but ultimately 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.

b) 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 experiment to give an indication of the 'direction of travel' from the interventions we tested rather than precise measures of the magnitude of any impact.

c) We were not able to test participants' reactions over time to the alert messages and changes to the prominence of the reporting function but "alert fatigue" is phenomenon which has been observed and documented in a number of different contexts e.g., healthcare[62].

d) Repeated exposure to these alerts could reduce their impact and cause users to ignore their content. Similarly, users could get used to reporting function to being more prominent. Testing the impact of exposure over time is something that can be tested in

[62] For example: Backman, R., Bayliss, S., Moore, D., & Litchfield, I. (2017), Clinical reminder alert fatigue in healthcare: a systematic literature review protocol using qualitative evidence, *Systematic reviews*, 6(1), 1-6

an experimental set-up (e.g., by repeating the experiment with the same participants) but may be better undertaken in collaboration with an online platform in a real-world setting.

e) The sample of participants we used in the research was set up to be nationally representative, but it was not large enough to pick up variations in viewing or reporting behaviour between different users e.g., different age groups. We are aware from qualitative research carried out by Ofcom and RCTs on platforms such as TikTok (see Annex 1) that there can be differences in online behaviour between older and younger generations. Again, this is something that would be better researched via a field RCT in collaboration with an online platform.

4.52 We would welcome the opportunity to explore the scope for conducting RCTS in collaboration with industry stakeholders.

A1. Annex: users' experiences of online safety measures

A1.1 In this section we give a brief overview of the available consumer survey evidence about users' experience of online harms and their knowledge and understanding of online safety measures, including alert messages and reporting mechanisms. We then briefly summarise some of the academic research we have considered in relation to alert messages and reporting mechanisms. Finally, we set out some examples of how alert messages and reporting mechanisms are presented to users of VSPs and online platforms.

Ofcom's Consumer research

Experience of harmful content across VSPs is relatively low

A1.2 Ofcom research shows that a 62% of internet users aged 13+ have come across potentially harmful content or behaviour online in the last four weeks.[63] Of those, 8% said that they experienced this on 'a website or app where you view videos posted by other users, e.g., YouTube TikTok'.[64]

A1.3 The research found that users aged 18-34, those from minority ethnic backgrounds, users from AB socio-economic groups, and those with a limiting or impacting condition are more likely than average to have experienced potentially harmful content or behaviour in the last four weeks.[65]

Less than half of VSP users report potentially harmful content

A1.4 Ofcom's Online Nation report shows that, of the 8% of users who said they experienced their most recent harm on a video sharing site or app, 36% flagged or reported the content.[66] This compares to an average of 31% for potential harms encountered on any type of platform.

A1.5 Looking at flagging or reporting across all platforms, a number of groups of users were more likely than average to have flagged or reported their most recently experienced potential harm. These include: Muslim users (44%), gay or lesbian users (41%), users from a minority ethnic background (37%), 25-34s (35%), parents (35%), and those with an impacting or limiting condition (34%). White users and those aged 13-17 were less likely than average to have done the same.[67]

---

[63] Ofcom, Online Nation: 2022.
[64] We recognise that this definition of a VSP is relatively narrow.
[65] Users from the AB social groups: 67%. Users aged 18-34: 65%. Users from a minority ethnic group: 67%. Users with a limiting or impacting condition: 70%.
[66] Ofcom, Online Nation: 2022.
[67] Figures relate to those who experienced their most recent harm on *any* type of platform.

Not knowing how to report an important factor in not reporting content

A1.6 Previous Ofcom research indicated that the majority of VSP users were unaware of the safety measures available on the VSPs they use. Although flagging and reporting tools were the most widely recognised safety tool (60%), they had only been used by one in four users. That research indicated that just over a third of VSP users (35%) did not take any action after being exposed to a potential harm. The reasons given by respondents for not taking any form of action included a perception that it would not make a difference (31%); or that they were not directly impacted (30%). A quarter said they didn't know what to do or who to report it to.[68]

User attitudes to protection and responsibility are mixed.

A1.7 A significant minority of UK internet users, 43%, agreed with the statement 'It is the responsibility of the website or app to control what is posted on their site', compared to 23% who agreed it is 'the individual's responsibility to ensure what they are posting is appropriate for other users'. Over a third of users (34%) agreed with neither of these statements more than the other.

A1.8 Half of all UK users aged 13+ believe that more safety measures are needed online, compared to 23% who think current measures are sufficient but, again, a substantial minority of 28% gave a 'neutral' response which didn't favour either statement.

A1.9 Ofcom research also shows there is no clear majority opinion about where to strike the balance between free speech and user protection: 38% agreed that 'It is important to monitor and delete offensive views to protect other users', while 34% agreed with the opposing statement that 'The Internet has an important role in supporting free speech, even when some users might find the content offensive.' Again, the proportion of users who were either unable or unwilling to favour one statement over the other was relatively large at 27%.

Literature Review

Alert Mechanisms

A1.1 There appears to be little evidence in the public domain relating specifically to the impact of different types of alert messages or warning messages on users' propensity to view online content.

A1.2 We are aware of a number of pieces of research which have focused on the use of alert messages or prompts in relation to users posting content and there appear to us to be some useful crossovers from that research to our work on the effectiveness of alert

[68] Ofcom, Video sharing platforms usage and experience, 2021.

mechanisms, not least in discussing the relevant underlying cognitive limitations and biases.

A1.3 Wang et al (2014) carried out a six-week field trial on 28 Facebook users testing two different interventions aimed at reducing users' regret from posting content online[69]. In the first intervention, the researchers tested the impact of a prompt which combined visual reminders about the prospective audience for the post together with an indication of the size of the audience for the post. This nudge targeted issues to do with bounded rationality and asymmetric information. The second intervention introduced a 10 second delay (using an actual countdown to publishing the post). This nudge was drawn from the literature on hyperbolic discounting and encouraged the user to pause and think about the post they were about to publish. The researchers found that reminders about the audience for posts could help to prevent unintended disclosures without imposing a major burden. However, they found that introducing a time delay before publishing users' post could be perceived as both beneficial and annoying.

A1.4 More recently, Katsaros et al (2021) tested an intervention aimed at users about to post harmful content on Twitter using an RCT.[70] The intervention they tested involved creating a prompt which created the opportunity for the user to pause and reconsider their Tweet. They found that users in the intervention posted 6% fewer offensive Tweets than non-promoted users. They also found that as well as a reduction in the number of offensive Tweets created in the future, prompted users received fewer offensive replies to prompted Tweets. They concluded that interventions allowing users to pause and reconsider their comments could be an effective mechanism for reducing offensive content online.

A1.5 TikTok worked with the consultancy Irrational Labs to test an intervention aimed at reducing the spread of misinformation on TikTok.[71] Again using an RCT, Irrational Labs designed a pair of prompts that were put on videos with "unsubstantiated content" i.e., content that fact-checkers had not been able to verify.

A1.6 The researchers again targeted the idea that the prompt would require users to pause and consider before viewing or sharing the content. They explicitly made the distinction between users having "hot states" where they act on emotions and "cold states" where actions are more logical and deliberate. They described TikTok is a "fast platform where users often act in hot states".

A1.7 The intervention had the effect of reducing shares of this content by 24% compared to the control group. In addition, the intervention also reduced likes by 7%, and views by 5%. The researchers also found that having a prompt or delayed prompt (after 3 seconds) didn't make any difference to viewing or sharing video content. As a result of this testing and trialling it was reported that TikTok was going to roll out this prompt across its platform.

[69] Wang et al (2014) "A field trial of privacy nudges for Facebook".
[70] Katsaros et al (2021) Reconsidering Tweets: Intervening during Tweet creation decreases offensive content.
[71] Irrational Labs How behavioral science reduced the spread of misinformation on TikTok.

Reporting Mechanisms

A1.8 There also appears to be very limited research regarding users complaining about or reporting content and their use of reporting mechanisms.

A1.9 There is a significant literature regarding consumer complaints behaviour. However, this focuses predominantly on customers' behaviour in making complaints about traditional goods and service industries in an offline world. Seminal work by Singh (1988) and Singh & Wilkes (1996) built taxonomies for consumer complaints behaviour and developed a conceptual framework for when and how consumers complain.[72] [73] It's not clear how applicable this research is to the issue of online content reporting mechanisms.

A1.10 The literature concerning online complaints behaviour relates to public reviews or complaints, and the effect on the complainant and other consumers. Work by Marx & Nimmermann (2017) investigates the handling of public complaints online and define response strategies that are beneficial for companies, while Sparks & Browning (2010) looked at the influential effect of public reviews on the hotel industry.[74] [75] Again, it is not clear how applicable this research is to online content reporting mechanisms.

Observational research

Alert Messages

A1.11 A number of online platforms already make use of alert or warning messages in specific circumstances. In some cases, the messages are focused on warning the users about the contents of the video they are about to watch. [76] In other instances, the alert messages apply when users are about to share content[77] or comment on content.[78]

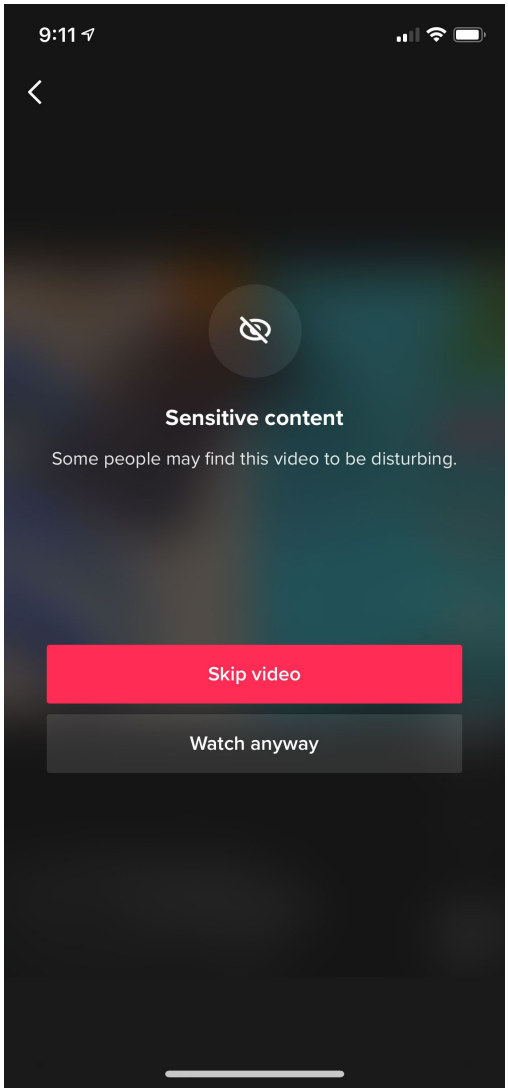
A1.12 The precise wording of the alert messages varies and in a number of instances the user is also presented the default setting to proceed to watch the video.

A1.13 In first screenshot below (Figure A1), we see an example of a user on TikTok being presented with a generic warning about the content that they are about to watch i.e., the warning is that "Sensitive Content: Some people may find this video content to be disturbing". We draw attention to the fact that screen is blurred out altogether and that the default option is set to skip the video (the bright red button).

---

[72] Singh (1988), Consumer Complaint Intentions and Behavior: Definitional and Taxonomical Issues.
[73] Singh & Wilkes (1996), When Consumers Complain: A Path Analysis of the Key Antecedents of Consumer Complaint Response Estimates.
[74] Marx & Nimmermann (2017), Online Complaints in the Eye of The Beholder: Optimal Handling of Public Consumer Complaints on the Internet.
[75] Sparks & Browning (2010), Complaining in Cyberspace: The Motives and Forms of Hotel Guests' Complaints Online.
[76] The Verge (2017), Instagram will begin blurring 'sensitive' posts before you can view them.
[77] The Verge (2020), Twitter is bringing its 'read before you retweet' prompt to all users.
[78] The Guardian (2019), Instagram's anti-bullying AI asks users: 'Are you sure you want to post this?'.

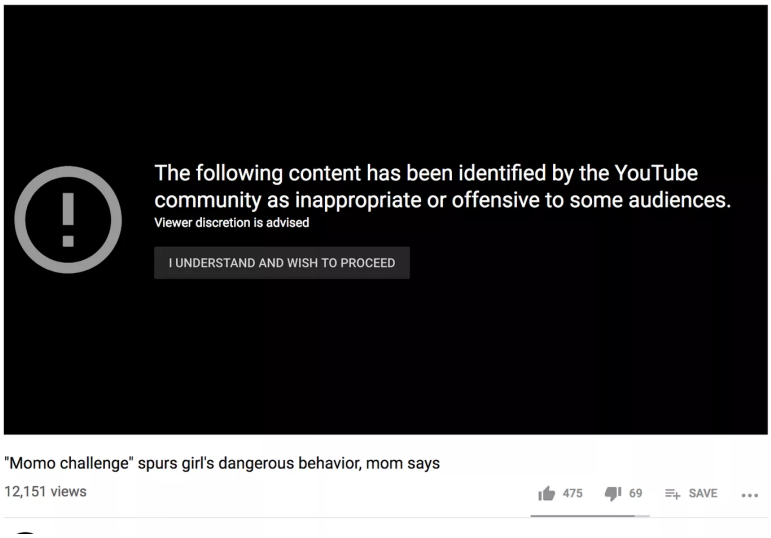
Figure A1: Example of a generic alert message on TikTok.[79]



A1.14 In another example (Figure A2), a user on YouTube is presented with an alert message which contains a social proof message i.e., the alert message refers to the fact that the content has been identified by the "YouTube Community" as inappropriate or offensive to some users. Again, the video itself is obscured (but there is a description at the bottom of the screen) and in this case the default option available to the user is "I understand and wish to proceed".

[79] TikTok (2020), Refreshing our policies to support community well-being.

Figure A2: Example of "social proof" alert messages found on YouTube



A1.15 In the example below (Figure A3), Twitter uses a warning message that is specifically about misinformation. As can be seen the warning label sets out the fact that some or all of the content in the Tweet is disputed and may be misleading. The option presented to the user is to "Learn More".

Figure A3: Example of a warning about misinformation on Twitter

Reporting Mechanisms

A1.16 How users access the reporting mechanism differs across platforms. In order to find the reporting mechanism, some platforms display a 'Flag' icon which is consistently visible to the user, whereas others locate the reporting mechanism within a drop-down menu, typical behind an ellipsis icon.

Figure A4: Example of the use of a reporting 'flag icon' by Vimeo.[80]

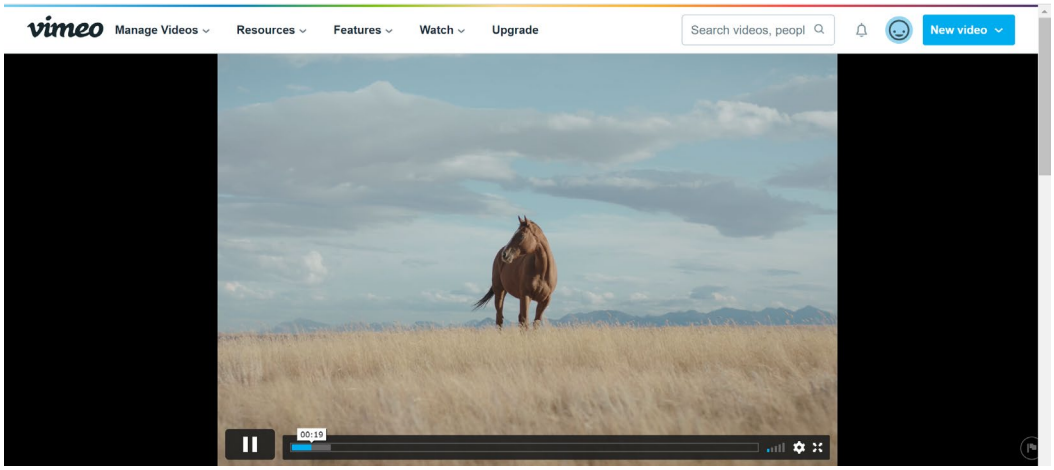
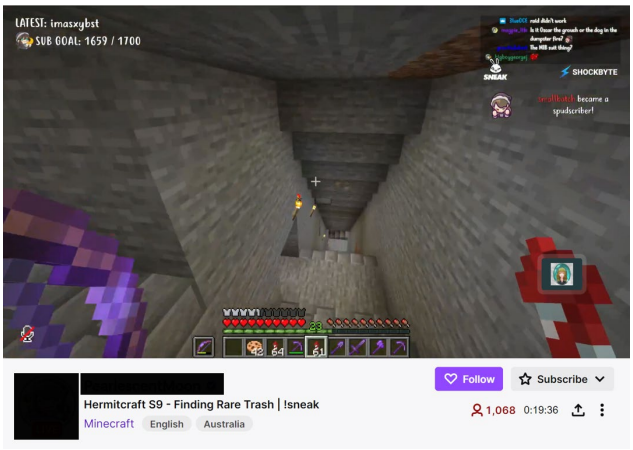


Figure A5: Examples of the report function sitting behind the '…' option on Facebook and Twitch.[81]



A1.17 Once into the reporting mechanism, the number of stages, the category labels, and the level of detail required by the user (usually in the form of a free text box) vary significantly. It also appears that the level and consistency of feedback to reports varies across platforms.

[80] Screenshot captured on 20 May 2022.
[81] Screenshots captured on 20 May 2022.