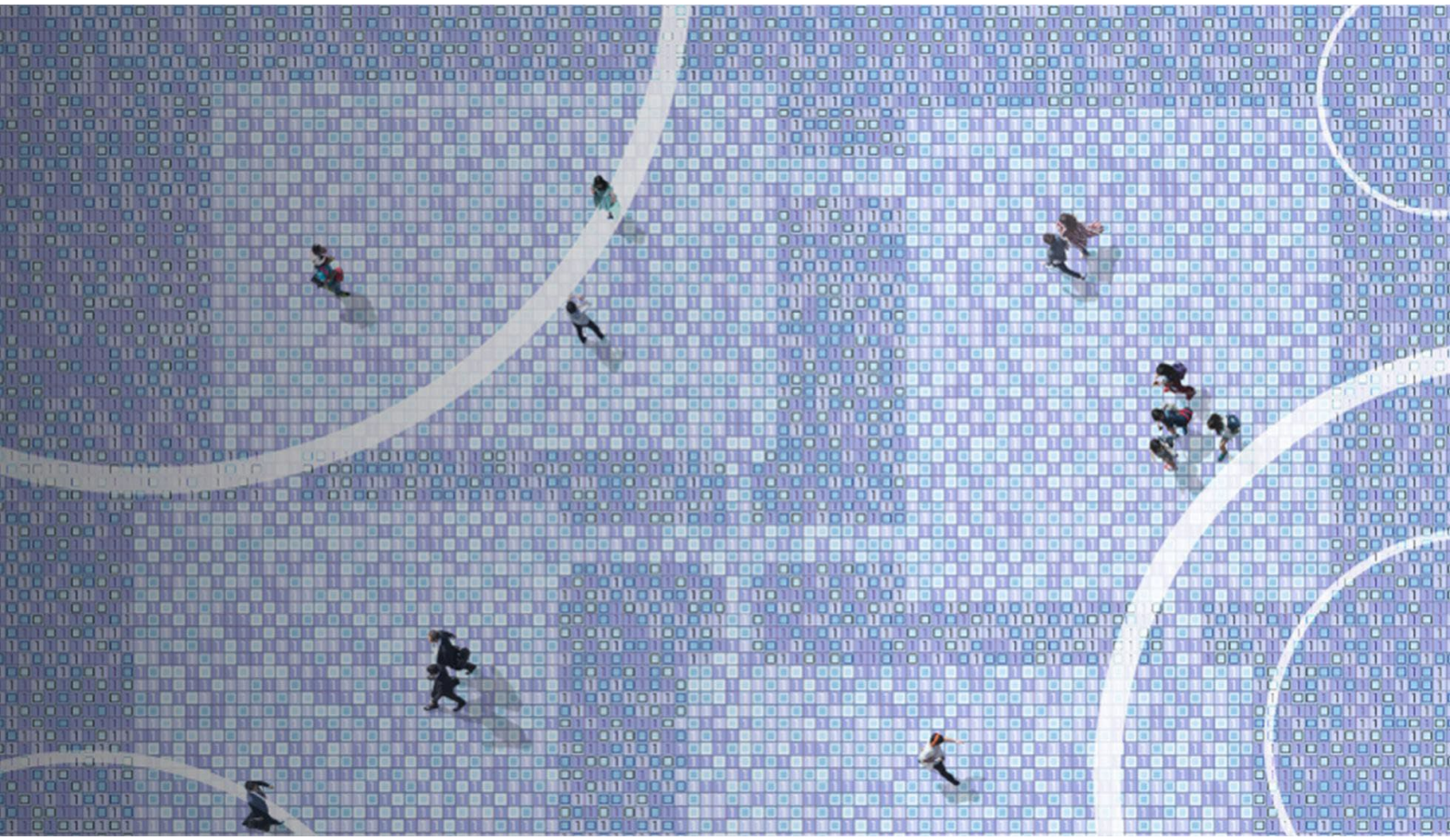
KANTAR PUBLIC

Ofcom online trials: safety features of video sharing platforms

Report

Kantar Public Behavioural Practice: Michael Ratajczak, Rupert Riddle, Yuchen Yang, and Natalie Gold

Ofcom: Alex Jenkins, Jonathan Porter, Amy Hume, Pinelopi Skotida, and Rupert Gill



1. Background and objectives

1.1. Regulatory Context

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.

In addition, as of November 2020, Ofcom oversees the regulatory regime which requires UK-established Video Sharing Platforms (VSPs) providers to include measures and processes in their services that protect users from the risk of viewing harmful content. Measures taken by a provider must be appropriate for the purposes of protecting users and must be effective in achieving this purpose. However, there is limited research in the public domain about their effectiveness.

VSPs—and social media in general—have the capacity to bring an extremely wide range of content direct to any user in a way that encourages immersive engagement. In many cases, this immersive engagement with different types of content will have positive effects (for example, discovering related information after watching an educational video).

However, in some cases the content could be illegal, and users should not be exposed to it. Alternatively, the content could be legal but carries with it the risk of causing psychological, physical, or financial harm, to particular groups of individuals. It is important that users have the ability to make an informed decision about whether or not to watch such content (in conjunction with other safeguards that platforms use to safeguard their users e.g., reporting mechanisms, parental controls and terms and conditions).

Ofcom is looking to research different methods for testing the effectiveness of different safety measures used by online platforms to safeguard users from harm more broadly and looking to build an evidence base about the effectiveness of specific techniques or interventions.

1.2. Experiment aims and objectives

Defaults are a very powerful form of a 'nudge' and they are one of the most well evidenced and robust forms of behavioural intervention. Defaults have been found to have a significant impact on user behaviour in a range of different settings across a range of products and services.

In the context of online safety, we are aware that a number of platforms already employ an "auto-play" default whereby a new video will start to load up as soon as a previous video clip comes to an end.

In the case of this research, the focus was on investigating the impact of alert messages used in conjunction with a default to auto-play or auto-skip content after the alert message has been on screen for a short period of time in the context of helping users make more informed decisions about whether to watch potentially harmful content. Specifically, it was of primary interest to investigate what impact having the default auto-play or auto-skip function has on the probability of skipping potentially harmful video content, compared to a control arm with no alert messages. It was also of interest to Ofcom to test what impact these defaults have on the probability of skipping potentially harmful video content, compared to an arm in which users had to make an active choice to skip or play content.

1.3. Research questions

In this trial, we aimed to answer the following three research questions:

RQ.1. What is the effect of having an alert message which also include either an auto-play or auto-skip functions compared to when there is no alert message?

RQ.2. What is the effect of having an alert message which also includes either an auto-play or auto-skip functions compared to when there is an alert message that requires an active choice to skip or play?

RQ.3. What is the difference in skipping behaviour between the auto-skip alerts and auto-play alerts?

2. Sample and data collection

2.1. Sample

The target population, in this study, consisted of adult UK VSP users. This experiment aimed to provide a sample that was as representative as possible, with respect to key demographic characteristics, of the UK's VSP users. Consequently, demographic quotas based on the adjusted quotas used in previous research with Ofcom were set. Specifically, the quotas in this trial were based on the relative proportions of respondents in each demographic sub-group who used VSPs at least once in the 12 months prior to participating in the first online behavioural Randomised Controlled Trial (RCT) conducted for Ofcom in relation to reporting mechanisms ("Ofcom 1").[1] The same demographic quotas were used in the second online behavioural RCT conducted for Ofcom in relation to the effectiveness of alert messages ("Ofcom 2").[2] Critically, participants who participated in the Ofcom 1 and Ofcom 2 trial were not allowed to participate in this trial.

A total of 2,801 UK participants, **aged between 18 and 69,** were recruited from Kantar's Lifepoints panel. All participants were asked whether they had used a VSP in the past 12 months in response to a screener question provided at the beginning of the experiment. Participants who had not used a VSP in the past 12 months were excluded from participating in this trial.

Kantar Public conducted this experiment online, using a device-agnostic platform; as such, participants were able to complete the experiment on a computer, mobile, or tablet, subject to participants' preference. Fieldwork took place across February-March 2023 over a four-week period.

Table 1 shows the quotas set before the recruitment began, and the quotas that were met when recruitment ended.[3]

Table 1. Demographic parallel[4] quotas set at the start of the study, and the quotas achieved

Demographics		Start	Finish
Gender	Male	49%	49%
	Female	51%	51%
	Other	-	<1%
	Prefer not to say	-	<1%
Age	18-24	14%	14%
	25-39	34%	34%
	40-54	30%	30%
	55-69	22%	22%
Ethnicity	White	87%	86%
	Mixed/Multiple Ethnic Groups	2%	2%
	Asian/Asian British	7%	6%
	Black/African/Caribbean/ British	3%	3%
	Other Ethnic Group	1%	1%
	Prefer not to say[5]	-	2%
Socio-economic grade	ABC1	56%	56%
	C2DE	44%	44%
Country	England	84%	84%
	Wales	5%	5%
	Scotland	8%	8%
	Northern Ireland	3%	3%

[1] Ofcom Economic Discussion Paper: "Understanding the impact of VSP design on user behaviour"
[2] Op.cit.
[3] Note that recruitment had to be restarted, because the panel provider set the wrong quotas on countries. Consequently, we over-sampled with respect to Wales and under-sampled with respect to Scotland. To rectify this, 79 participants from across all trial arms who lived in Wales when they completed the survey were removed (at random from each arm), and 79 participants who lived in Scotland at the time when they completed the survey were recruited (and randomly allocated across the trial arms).
[4] When using parallel quotas, the sample will aim to fulfil all required quotas on age, gender, SEG, location and ethnicity. However, those proportions would not be interlocked with each other. This would mean a final sample with the correct proportion of each category, i.e., 49% male, 51% female, 56% SEG ABC1, 44% SEG C2DE etc. Although it is theoretically possible that all the male participants might end up in SEG ABC1, we did not see any examples of this issue arising.
[5] Includes participants who did not agree to be asked this question (n = 34) and those who refused to answer this question when asked (n = 9).

2.2. Data collection

Kantar Public ensured compliance with the Data Protection requirements in the UK, including the UK's General Data Protection Regulation (UK GDPR).

In addition, participants were able to opt out of the study. Participants were notified, at the beginning of the study, that they may be exposed to what they could consider to be harmful videos and informed consent was obtained for the collection of sensitive data, such as ethnicity, from the respondents.

The consent and potentially harmful videos were reviewed by Kantar's Profiles' Privacy team and Kantar Public's Global Head of Quality, Information and Security. In addition, this team assessed and documented what data would be collected, how it would be collected, and determined that the data would be collected in compliance with Profiles' data protection framework.

2.3. Randomisation

Participants were randomly allocated into one of the experiment's four arms, three of which included interface-based interventions that aimed to help participants make a more informed choice about viewing potentially harmful video content.

To allocate respondents to experimental arms, a method of blocked randomisation was used (least-filled quotas). This method ensured that blocks filled at a consistent rate whatever the sample size. Note that this method of randomisation is frequently used in behavioural economics related studies,[6] as well as in clinical trials,[7] and was successfully used to recruit participants in the previous two online behavioural RCTs for Ofcom.

2.4. Incentivisation

Panel participants were paid a sum of '*Lifepoints'* for completing the experiment. These points can be exchanged for vouchers or cash via PayPal.

2.5. Ethics

The purpose of the experimental environment was to replicate the real-world context as closely as possible, to get as close as possible to actual VSP users' behaviour. It would have been difficult, if not impossible, to gain externally valid evidence of the propensity to skip potentially harmful content in an experiment that did not expose participants to potentially harmful content. However, exposing participants to content that could be deemed 'very harmful' would not have been ethically acceptable, and to attempt to do so without mitigation would present a high risk to participants.

Kantar Public's Behavioural Practice team reused some of the videos from the two previous behavioural experiments with Ofcom. However, two videos were replaced, namely the Neutral Two ('Blue Origin Booster Landing') and Three ('Celebrity Breakups') videos, with 'Optical Illusion' (Neutral One) and 'How to do a pull-up' (Neutral Two) to ensure that the material was relevant and engaging. The potentially 'harmful' video content (content that some participants could consider to be harmful) used in Ofcom 1 and Ofcom 2 was selected for inclusion by:

1. Searching various VSPs for legal but potentially harmful videos that:
    - had been made downloadable by their originators so they could be downloaded directly from the website
    - were not deemed as too extreme to include in the experiment
    - were thought to hold viewer's attention in the first 20-45 seconds
    - were considered to be relevant.
2. Sharing these videos between the Kantar Public's project team and Kantar's Profiles' Privacy team to confirm that these videos could be considered as harmful by some participants, but that these videos were, nonetheless legal and acceptable for provision to participants.

This type of content, while still potentially harmful to some participants, was more acceptable for inclusion because of the content's lower impact and greater prevalence (and hence likelihood of being seen 'for real'). Ofcom's own research indicates that 62% of internet users have experienced at least one instance of potentially harmful behaviour or content online in the last four weeks.[8]

The following steps were taken to mitigate any residual risk in the experiment:

- An upfront consent screen at the start of the experiment informed participants that they would be shown some content that could be considered harmful; participants were allowed to refuse to participate if they did not want to be exposed to this.
- A debrief screen at the end of the experiment provided web links to support on any of the potential harms included in the content shown in the experiment.

[6] Dannenberg, A., & Martinsson, P. (2021). Responsibility and prosocial behavior-Experimental evidence on charitable donations by individuals and group representatives. *Journal of Behavioral and Experimental Economics, 90*, 101643.
[7] For example: https://onlinelibrary.wiley.com/doi/full/10.5694/j.1326-5377.2002.tb04955.x
[8] Ofcom. (2022). *The Online Experiences Tracker (2021/22): Summary Report*. https://www.ofcom.org.uk/__data/assets/pdf_file/0025/244168/online-experiences-tracker-waves-1-and-2-summary-report.pdf

In addition to the above, in the study, participants were able to skip any of the video content at any point. This meant that they were not required to watch any of the videos if they did not want to.

2.5. Disclaimer

Ofcom did not have any role in either identifying, reviewing, or selecting the potentially harmful video content selected by Kantar Public for use in this study. Kantar's Profiles' Privacy team ensured that the research process complied with the relevant regulations, such as the UK GDPR, and best practice (see also section 2.2). Kantar Public also adhered to the Market Research Code of Conduct 2019.

2.6. Attention tests

To keep the quality of data high and remove any skimmers who were attempting to get through the experiment as quickly as possible, two attention checks were included in this experiment.

First, any respondent, who completed the study in less than 40% of the median completion time for all respondents, was removed. Second, any respondent who failed to correctly answer the attention check question was excluded from the study. The attention check specified: "Please select the ''Purple" option below. We are asking this for quality control reasons to check you are paying attention to the questions in the survey."

The response options were:

Blue	**Orange**	**Green**	**Red**	**Pink**	**Purple**	**Brown**

The total drop-out rate, due to responders completing the study too quickly or failing the attention check, over the whole study, was 9% (116 panellists failed the attention check question; 162 panellists were excluded based on completing the experiment too quickly).

2.7. Soft launch

To ensure that there were no unforeseen issues with the experimental design and script, an initial soft launch involving 10% of participants was conducted in February. During the soft launch the following was monitored: the drop-off rate, time to finish the experiment, view time of each of the videos, the quotas, and whether we captured the required data. We identified one data capture issue, but this did not have an impact on our ability to analyse the data.[9]

[9] We found one error in data capture of the skip functionality. Specifically, there was an issue with recording of skipping behaviour in Arm 3 (alert messages with auto-skips) for one of the potentially harmful videos (Video 5: Covid-19 Misinformation). Specifically, auto-skips at the alert stage were not recorded as alert stage skips. However, we were able to correct this error by re-coding alert stage skips to include auto-skips.

3. Trial design and flow

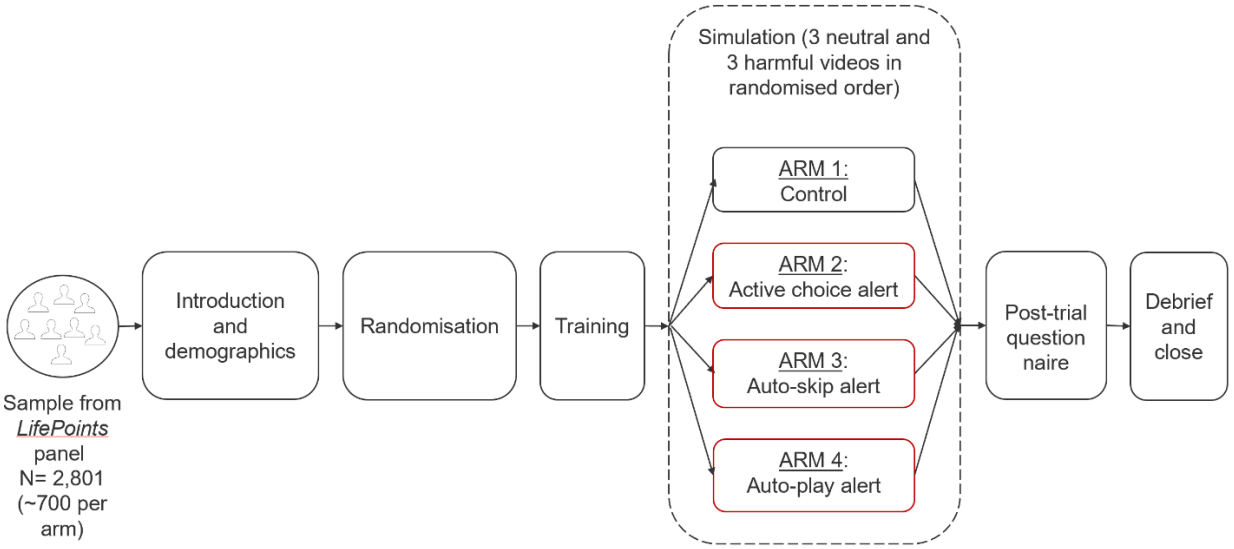


Figure 1. Trial design and flow

3.1. Introduction and participant consent

Participants were first presented with an introduction screen thanking them for taking part in the study and outlining what participation in the study involved. The introduction screen contained a disclaimer about the inclusion of potentially harmful content that read *"Some of the videos you will see may contain violence, extreme views, or harmful content. If you do not wish to proceed, please opt out below."* and *"Are you happy to continue?".* An opt-out button "*No*" was provided at this point, and the number of participants who chose to opt out was recorded (Figure 2 shows participant flow).

There was also a debrief screen at the end of the experiment which provided a link (https://saferinternet.org.uk/report-harmful-content) to support on any of the potential harms included in the content shown in the experiment.

3.2. Demographics and VSP use screener

On entry to the trial, participants were asked demographic questions so that recruitment could be monitored against quotas of interest (age, gender, socioeconomic background, location, and ethnicity). Participants were screened for VSP use by asking which of 10 common video sharing platforms (Youtube, Facebook, Instagram, Snapchat, TikTok, Twitch, Onlyfans, Vimeo, Bitchute, Fruitlab) they had used within the past 12 months. Potential participants were screened out if they answered: "I haven't used any video sharing platforms in the past 12 months".

3.3. Training stage

Once participants confirmed their demographics, the interface that they would be using in the experiment was introduced. At this stage participants were randomly allocated to experimental blocks, and they had the opportunity to interact with the interface that they were allocated to. First, participants saw a static screenshot of the interface they were randomly allocated to with instructions for how they could use the buttons available and a short description of how the experiment would proceed.

The interface was a variation of a 'generic' VSP that had previously been found to increase the incidence of reporting of potentially harmful content (the salience intervention, Arm 2, in the Ofcom 1 trial), incorporating and varying features that are common to many platforms but without any familiar branding. After users saw the labelled screenshot, they were shown a training video that they were able to interact with by choosing to react (like/dislike[10]), comment or share (indicated by adding in comments or pressing the share button in the interface), report, or skip past to the next piece of content (see Figure 3).

Participants were able to 'play' with this version until they became familiar with how it worked. The video content showed to the participants at the training stage was selected in the same way as the videos for the main experiment part (see section 3.4). The video content for the training stage was unlikely to be classed as harmful by any participant, as it did not contain potentially harmful content

[10] Note that each video would already have a number of likes and views when participants saw the video. The counts of likes for videos were created using random generation for the Poisson distribution with n = 10,000. In other words, each video had approximately 10,000 likes. Views were generated in the same fashion, but n was 200,000. Overall, the number of likes was approximately 5% of the number of views.

("This is not a house": https://vimeo.com/555252697). After interacting with the training screen, participants were able to move on to the main experimental section.

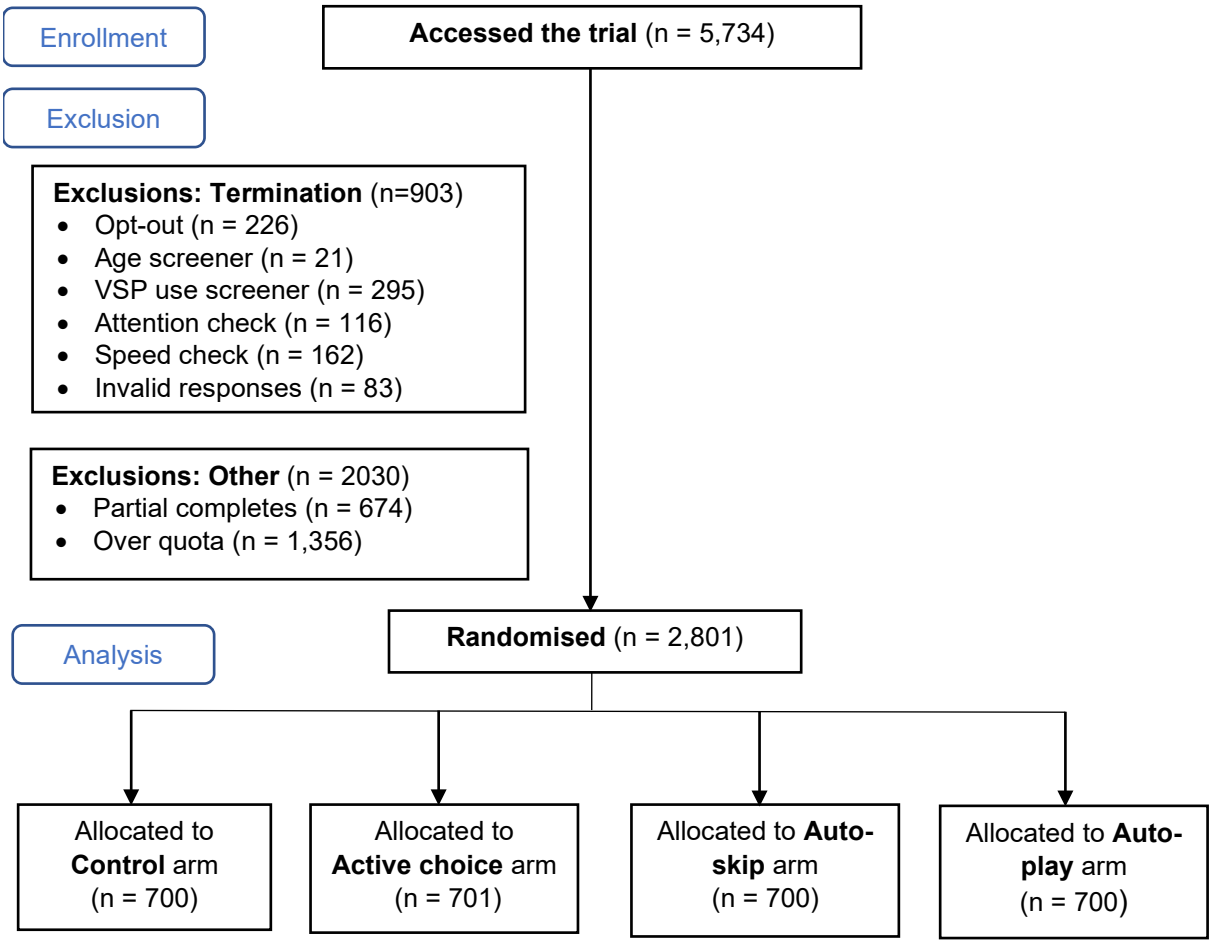


Figure 2. Participant flow diagram

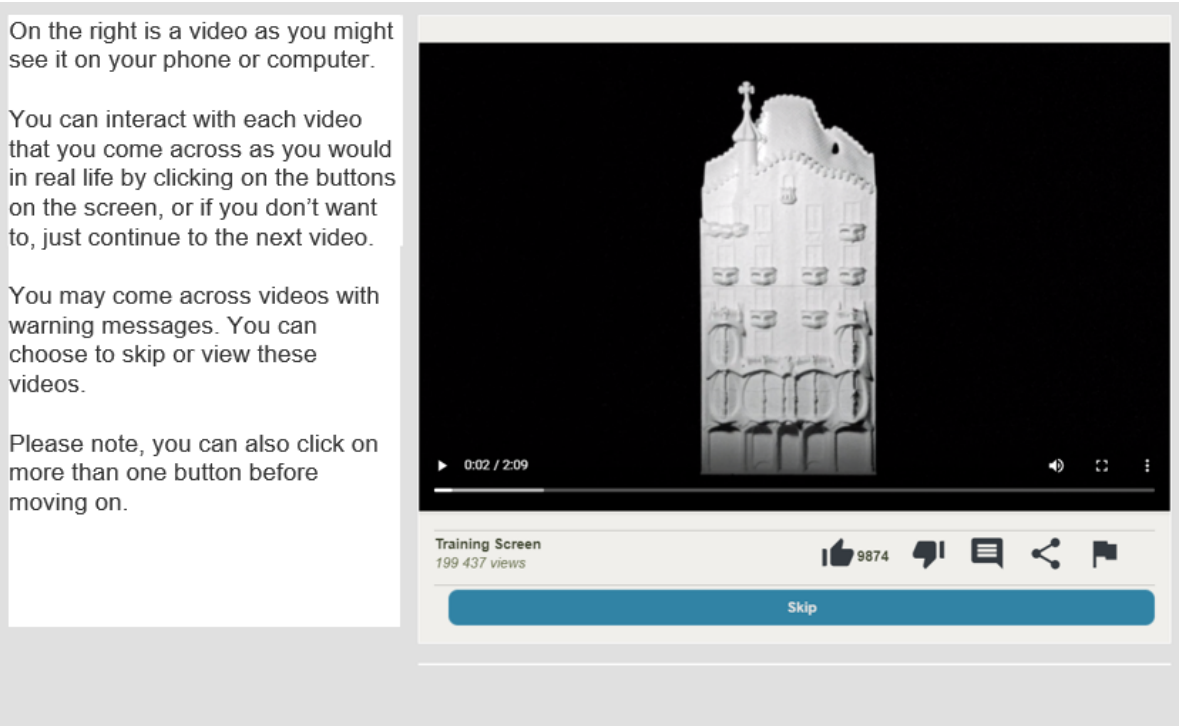


Figure 3. Control interface training screen

3.4. Main experiment

The experimental design aimed to enable users to make more informed decisions about whether to

watch video content.

In the main experiment, participants were exposed to six pieces of video content presented in a random order within the simulated VSP interface. Three pieces of content were neutral, and three were potentially harmful (of the type that was in the interest of the research outcomes to encourage people to make more informed decisions about).

Video content:

Neutral One: Optical Illusion: https://odysee.com/@AussieFighter:8/Optical-Illusion-chair:8

Neutral Two: How to do a pullup: https://odysee.com/@LiquidLoans:0/how-to-do-a-pull-up-tutorial:9

Neutral Three: Vegan Matcha Pancakes: https://vimeo.com/23873736[11]

Potentially Harmful One: Covid-19 Vaccine Misinformation (trimmed): https://vimeo.com/496630435

Potentially Harmful Two: Tube Racism Fight: https://leakreality.com/video/25086/repost-fight-breaks-out-after-british-man-racially-harass-asian-woman

Potentially Harmful Three: Homophobic / Offensive language (trimmed): https://leakreality.com/video/26960/uk-muslim-cleric-music-makes-you-gay

All videos were chosen, or trimmed, to be engaging in the first 20-45 seconds to hold participant attention and to ensure videos were not over a minute in duration. In addition, recent and relevant potentially harmful content was prioritised for the samereason.

3.5. Post-trial questionnaire

The participants were asked questions about the alert messages and their attitudes to the messages (see section 9. Appendix A: Post-trial questionnaire). The responses to these questions were used as secondary outcomes measures from the trial as well as assisting our understanding of trial outcomes.

[11] Hyperlinks to three videos (Neutral Three, Potentially Harmful Two, and Potentially Harmful Three) do not work (last checked on the of 22nd of March 2023). Since the work was conducted, Neutral Three: Vegan Matcha Pancakes has been deleted. In addition, https://leakreality.com has been taken down. Thus, it is not possible to provide working hyperlinks to these videos.

4. Interventions

There were four arms in this experiment, each outlined below and shown in section 4.2. These were developed and selected incollaboration with Ofcom:

1. *Arm 1 – Control (Figure 4)*: The control arm included an interface that was a generic version of a VSP,[12] without any alert messages informing users that they were about to see potentially harmful content.
2. *Arm 2 – Active choice alert (Figures 5 and 6)*: Participants saw "This video may contain sensitive material"[13] alert message before seeing each potentially harmful video. An active choice to skip or play was required in order to proceed to the next page. If the participant opted to skip the video, there would be a 500ms delay prior to the presenting next video.[14] The interface of the platform, other than for the alert message, was the same as in Arm 1 – Control.
3. *Arm 3 – Auto-skip alert (Figures 7 and 8)*: Participants saw the same alert message as those in Arm 2, but the alert message also included a countdown to auto-skip a potentially harmful video.[15] If participants did not engage with the alert message (i.e., didn't make an active choice), the video was skipped automatically after 5 seconds. If the participant opted to skip the video, or the video was skipped via the countdown timing out, there would be a 500ms delay prior to the presenting next video. The interface of the platform, other than for the auto-skip default was the same as in Arm 1 – Control and Arm 2 – Active choice alert.
4. *Arm 4 – Auto-play alert (Figures 9 and 10)*: Participants saw a variant of the same alert message as those in Arm 2, but the alert message also included a countdown to auto-play a potentially harmful video. If the participants did not engage with the alert (i.e., didn't make an active choice), the video was played automatically after 5 seconds. If the participant opted to skip the video, there would be a 500ms delay prior to the presenting next video. The interface of the platform, other than for the auto-play default, was the same as in Arm 1 – Control, Arm 2 – Active choice alert, and Arm 3 – Auto-skip alert.

4.1. Hypotheses

Hypothesis 1: The probability of skipping potentially harmful content would be significantly higher in an arm which included an alert message with a default auto-skip function compared to an arm in which no alert message is presented.

- The probability of skipping potentially harmful videos would be significantly higher in Arm 3 – Auto-skip alert compared to Arm 1 – Control.

Hypothesis 2. The probability of skipping potentially harmful content would be significantly different in an arm which included an alert message with a default auto-play function compared to an arm in which no alert message is presented.

- The probability of skipping potentially harmful videos would be different in Arm 4 – Auto-play alert compared to Arm 1 – Control.

Hypothesis 3: The probability of skipping potentially harmful content would be significantly higher in an arm which included an alert message with a default auto-skip function compared to an arm in which an alert message which requires an active choice to skip or play is presented.

- The probability of skipping potentially harmful videos would be significantly higher in Arm 3 – Auto-skip alert compared to Arm 2 – Active choice alert.

Hypothesis 4: The probability of skipping potentially harmful content would be significantly different in an arm which included an alert message with a default auto-play function compared to an arm in which an alert message which requires an active choice to skip or play is presented.

- The probability of skipping potentially harmful videos would be different in Arm 4 – Auto-play alert compared to Arm 2 – Active choice alert.

Hypothesis 5: The probability of skipping potentially harmful content would be significantly higher in

[12] The interface used was identical to that used in Arm 2 of the Ofcom 1 trial. The interface included a reporting icon on the main interface, but no additional prompts to report content, or any alert messages.
[13] After user testing of the trial, it was decided to move away from the trial protocol which used a "social proof" alert message and use a simpler generic alert message as to not overload the user with too much information, in combination with the countdowns and options to view or skip.
[14] The 500ms delay was introduced to Arms 2-4 to prevent accidental skipping, by double-clicking "Skip Video" at the alert stage, of more than one potentially harmful video.
[15] Originally, the countdown was below the options to "Skip Video" and "Play Video". However, the design was adapted after user testing. Specifically, the countdown was moved higher in the alert, between "This video may contain sensitive material" and the options to "Skip Video" and "Play Video", to minimise spacing (e.g., Figure 6).

an arm which included an alert message which requires an active choice to skip or play compared to an arm in which no alert message is presented

- The probability of skipping potentially harmful videos would be significantly higher in Arm 2 – Active choice alert compared to Arm 1 – Control.

4.2. Intervention designs

Figure 4. Arm 1 – Control (no alert message)

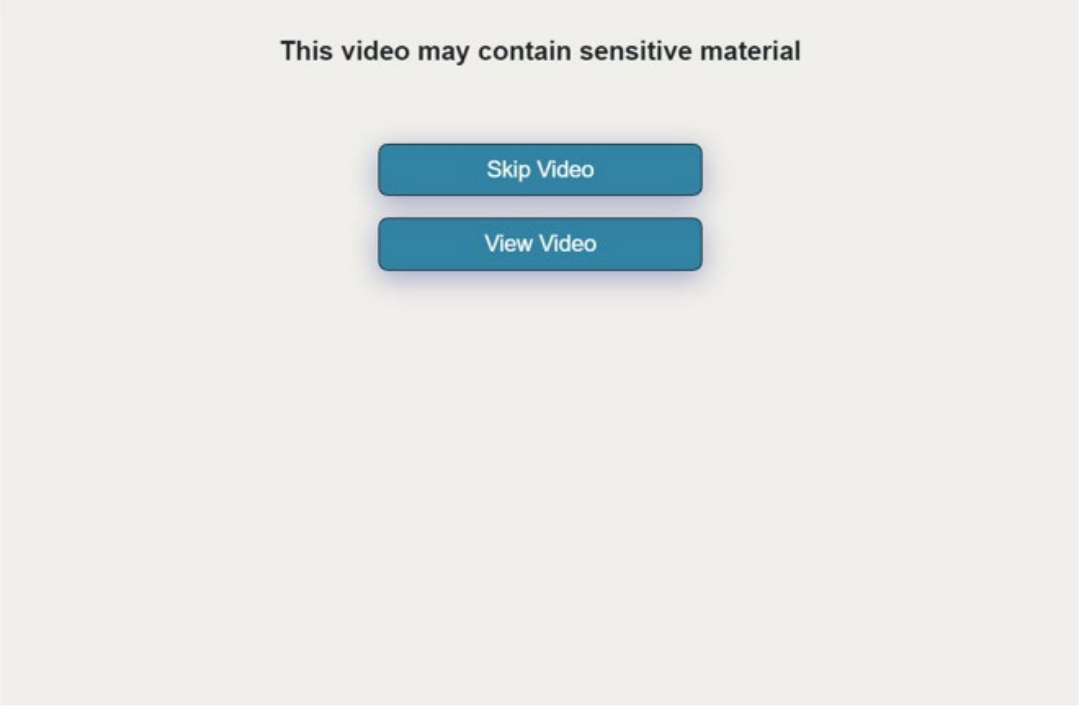


Figure 5. Arm 2 – Active choice alert

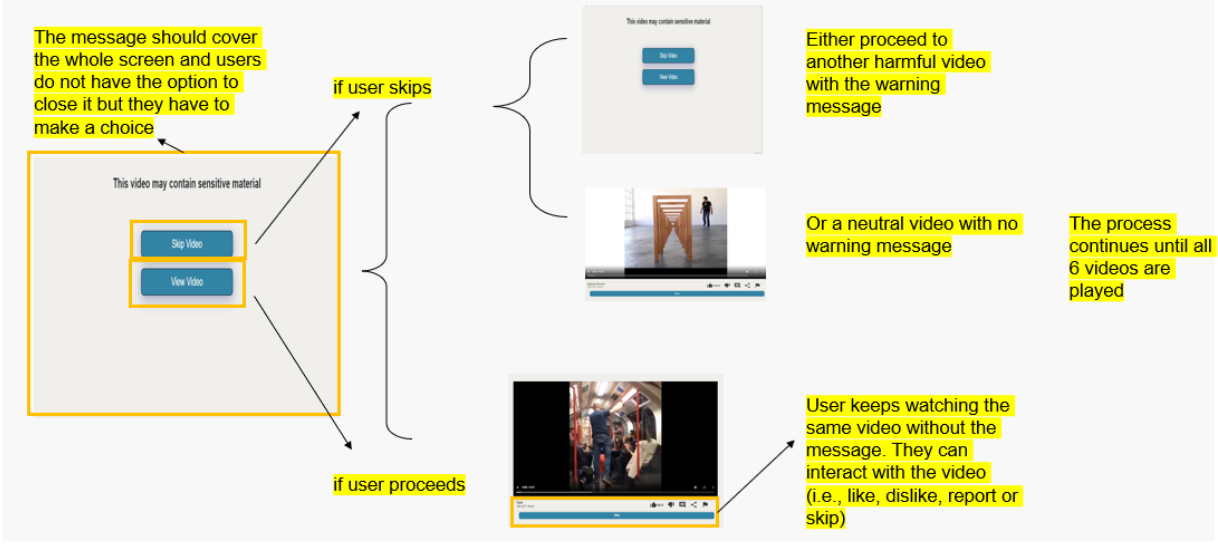


Figure 6. Arm 2 – Active choice alert (skipping flow)

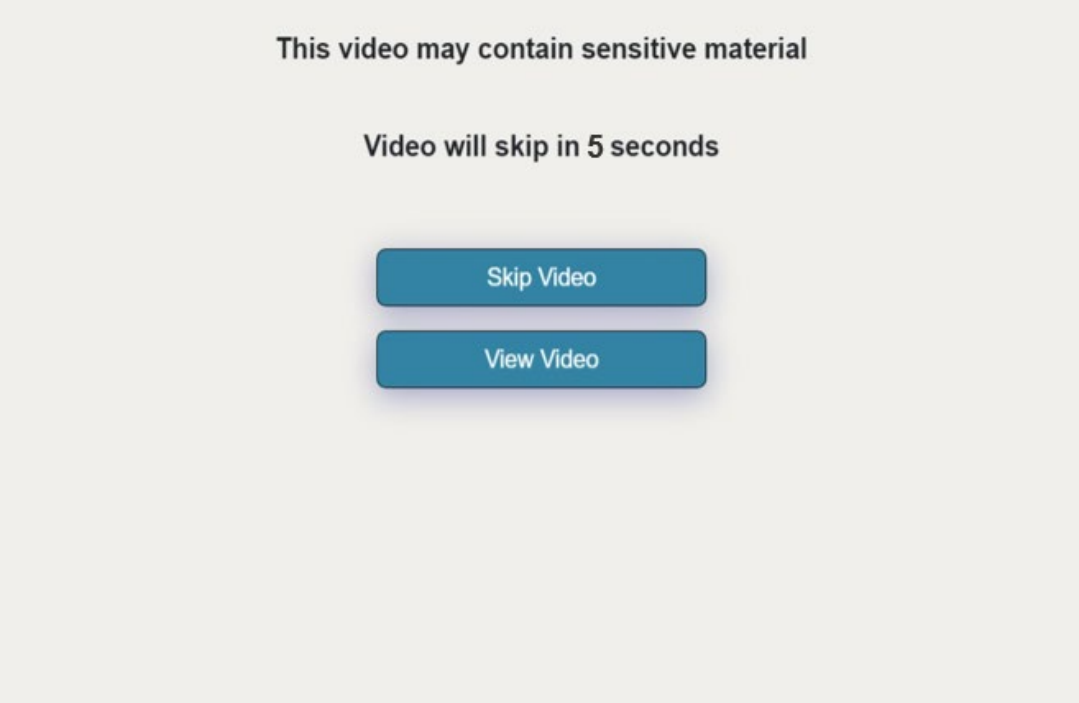


Figure 7. Arm 3 – Auto-skip alert (the countdown was dynamic)

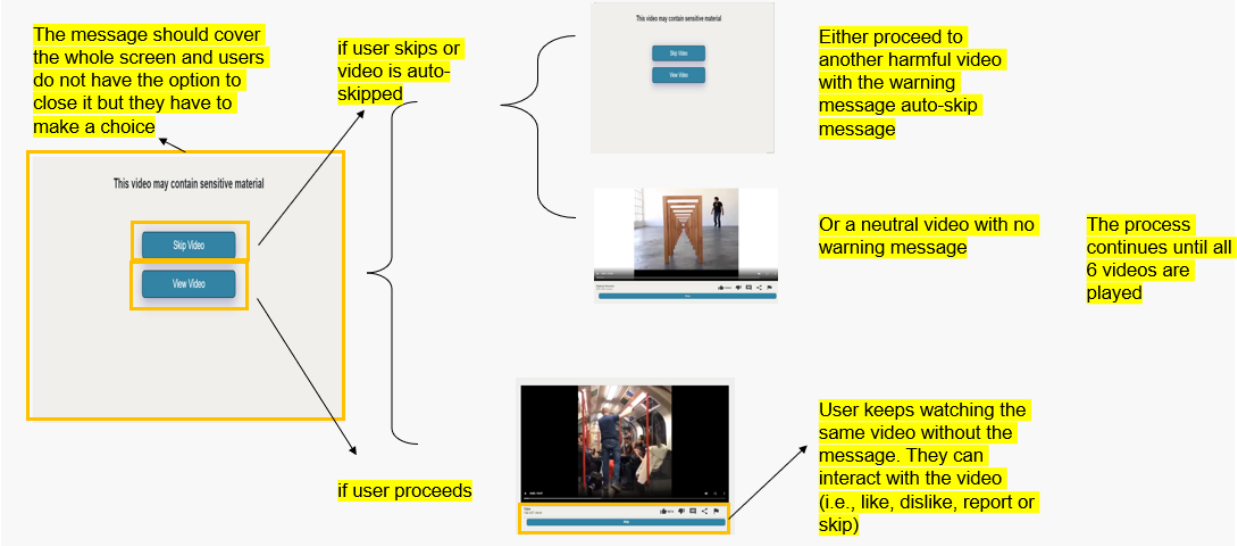


Figure 8. Arm 3 – Auto-skip alert (skipping flow)

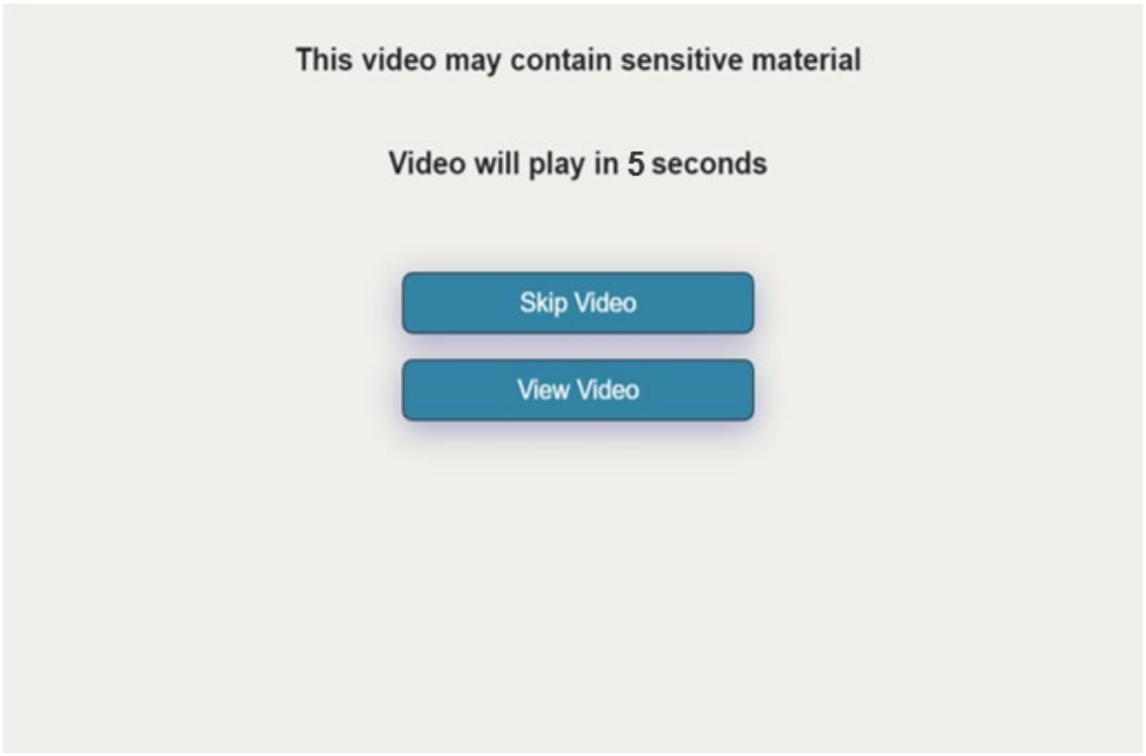


Figure 9. Arm 4 – Auto-play alert (the countdown was dynamic)

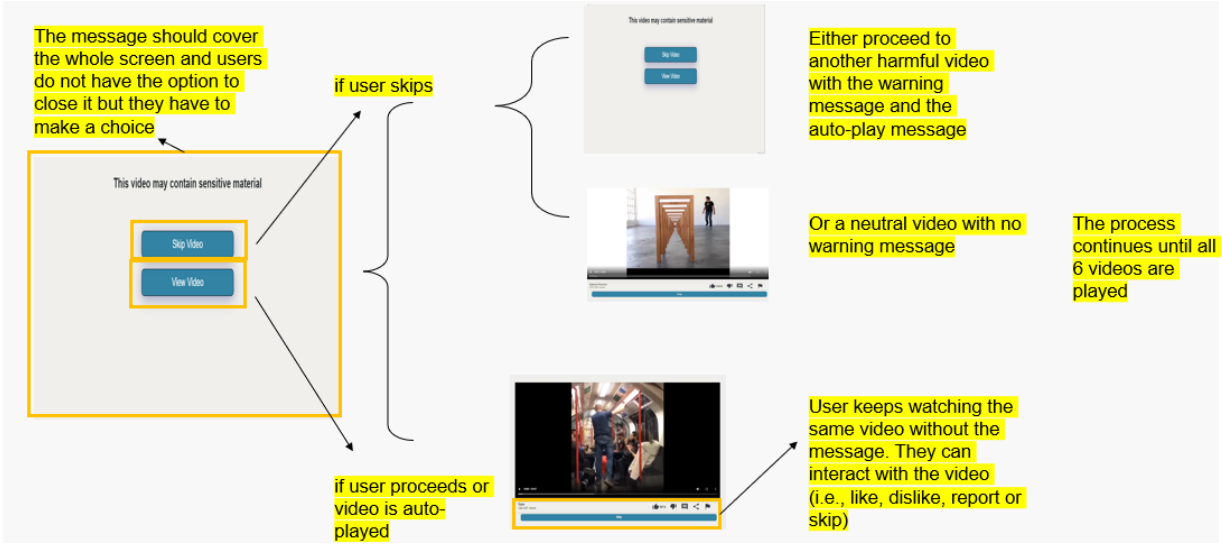


Figure 10. Arm 4 – Auto-play alert (skipping flow)

5. Outcomes

5.1. Primary outcome

In this study, we measured whether a participant had decided to skip (1) or not skip (0) each potentially harmful video (out of 3) at the alert messages stage or at the stage of watching the videos (if they decided to proceed). In the control and active choice alert arms, if a participant opted to skip a video at any point, their response was coded as 1, otherwise (if they watched the video all the way through) it was coded as 0.

In the auto-skip arm, if a participant did not make an active choice and the video was auto-skipped, this still counted as a skip (1). Likewise, in the auto-play arm, if a participant did not make an active choice and the video was auto-played until completion, this was recorded as not skipping (0) (note that if participants chose to skip once they had started to watch the video, the response was recorded as a skip (1)). This binary variable constituted the primary outcome variable (1.1 in Table 2).

5.2. Secondary outcomes

The first secondary outcome (2.1 in Table 2), whether a participant decided to skip (1) or not skip (0) each neutral video, was analysed in the same fashion as the primary outcome to examine whether alert messages had an impact on the probability of skipping of neutral videos.

The second secondary outcome was a binary variable indicating whether a user submitted a complete report of potentially harmful content when viewing each potentially harmful video (2.2 in Table 2). This outcome was measured to see how comparable the reporting behaviour of participants was compared to the reporting behaviour of participants in the Ofcom 1 trial. In addition, it allowed us to see whether there were any differences in the incidence of reporting between the different intervention arms in this study.

The third secondary outcome was the length of time participants viewed each potentially harmful video (2.3 in Table 2). Measuring this allowed us to examine whether participants in the intervention arms spent a different amount of time, compared to participants in the control arm, on watching potentially harmful videos.

The fourth secondary outcome was the time taken to respond to alert messages prior to viewing each potentially harmful video (2.4 in Table 2). Measuring this outcome allowed us to examine whether time to respond to alert messages changed with repeated exposure to alert messages over the course of the experiment.

Last, we captured responses to attitudinal questions at the end of the study (2.5 in Table 2). This allowed us to see whether there were any differences in self-reported attitudes to the alert messages between the intervention arms.

Table 2. The list of outcome measures and descriptive metrics used in the study

	Outcome measure
Primary	1.1. (Behavioural) Whether a participant decided to skip or not skip each potentially harmful video at either the alert message stage or at the stage of watching the potentially harmful video (if they decided to proceed)
Secondary	2.1. (Behavioural) A binary variable indicating whether a user skipped a neutral video that was binary coded as 1 if a user skipped a neutral video or 0 if they did not, at the video screen level (sensitivity analysis of the primary outcome) 2.2. (Behavioural) A binary variable indicating whether a user submitted a complete report of potentially harmful content when viewing each potentially harmful video 2.3. (Behavioural) The length of time participants viewed each potentially harmful video. If a participant skipped a potentially harmful video during the alert message stage, manually or via default, the length of time spent watching the video was recorded as NA 2.4. (Behavioural) The length of time to interact with the alert message (by either clicking “proceed” or “skip”) (length of time recorded to two decimal places) 2.5. Responses to attitudinal survey questions (section 9, Appendix A: Post-trial questionnaire) which ask whether participants thought that the alerts were

	useful, distracting, annoying, suggestive of appropriate action, or making them feel rushed. In addition, participants were asked whether they regretted watching potentially harmful videos
Descriptive metrics	3.1. Other forms of engagement: Number of likes, dislikes, shares, and comments a participant made on potentially harmful content and neutral videos 3.2. A binary variable for each video for each participant indicating whether a participant completed that action (1) or not (0) for all possible forms of engagement 3.3. A variable with aggregated count per participant of the number of interactions per participant 3.4. A binary variable that was coded as 1 if a user skipped the video at the alert message warning stage, and 0 if the user skipped the video after the alert message (at the interface stage). If the video was not skipped at any point, the variable was recorded as NA 3.5. A binary variable that was coded as 1 if a user actively skipped the video at the alert message warning stage, and 0 if the user passively skipped the video by waiting for the auto-skip (Arm 3 Auto-skip) 3.6. A binary variable that was coded as 1 if a user actively played the video at the alert message warning stage, and 0 if the user passively played the video by waiting for the auto-play (Arm 4 Auto-play) 3.7. The length of time from entering the interface to pressing any button, recorded to two decimal places 3.8. The length of time from starting a report (pressing the report button) to submitting a report (pressing the submit button) (to two decimal places) 3.9. The order in which participants were shown each video, irrespective of if the video was played or skipped. 3.10. The length of time participants viewed harmful and neutral videos, per video (to two decimal places). If a participant skipped a potentially harmful video during the alert message stage, manually or via default, the length of time spent watching the video was recorded as NA 3.11. A binary variable indicating whether a participant was shown 3 potentially harmful videos in a row or not (this was expected to be a relatively rare occurrence). 3.12. A binary variable which indicated whether a participant ‘opted out of the experiment, in order to determine the relative opt-out rate of participants across the entire experiment. 3.13. Number of participants who decided not to continue at the introduction stage 3.13. Number of participants who dropped out during the study 3.14. Number of participants who failed the attention check

6. Statistical methods and analysis

6.1. Statistical methods

Primary analysis

The primary outcome for the analysis was whether a participant decided to skip or not skip each potentially harmful video at the alert messages stage or at the stage of watching the potentially harmful video (if they decided to proceed) (see 1.1 in Table 2).

Given that this outcome was binary (skip vs. not skip) a logistic mixed-effects model was used to examine the differences between the different intervention arms. One of the key advantages of using this model was that it considered additional uncertainty due to the effect of variation in individual responses and due to the effect of variation in the potentially harmful video content. Additional motivation for including a random intercept for each video was that the skipping behaviour was expected to be different between different videos. In other words, it was not assumed that every potentially harmful video would have equal probability of skipping. Instead, it was assumed that some of these videos may have a lower or higher probability of skipping than others. Thus, a basic proposed model specification was:

$$Y_{ij} \sim Bernoulli\left(Y_{ij}^{0}\right), Y_{ij} \in \{0,1\}, Y_{ij}^{0} = Prob(Y_{ij} = 1)$$

$$Logit\left(Y_{ij}^{0}\right) = \beta_0 + \beta_1 Arm_i^2 + \beta_2 Arm_i^3 + \beta_3 Arm_i^4 + u_{1i} + u_{2j}.$$

In the equation above, Y_{ij} was binary variable indicating whether a participant i watching potentially harmful video j pressed the skip button or not. The binary variable was 1 if the video was skipped, but 0 if the video was not skipped.

β_0 was the predicted value for a baseline category - here Arm 1 – Control - whereas β_1, β_2, and β_3, represented deviations in the log-odds of skipping potentially harmful videos of Arms 2 – Active choice alert, 3 – Auto-skip alert, and 4 – Auto-play alert, respectively, from Arm 1 – Control.

u_{1i} was the random intercept of participant i, $u_{1i} \sim N(0, \sigma_1)$ for $i \in \{1, 2, \dots, N\}$ where N was the number of participants, and u_{2j} was the random intercept of potentially harmful video j, $u_{2j} \sim N(0, \sigma_2)$ for $j \in \{1, 2, 3\}$.

To answer our research questions (see section 1.3), using this model, the following hypothesis was tested:

$H_0^1: \beta_0 = \beta_k; H_1^1: \beta_0 \neq \beta_k \; where \; k \in \{1, 2, 3\}$.

In addition, to answer our second and third research questions (see section 1.3), comparisons between arms were performed. When running these multiple comparisons, the Bonferroni correction was utilised to maintain the family-wise error rate.

Sensitivity analyses of the primary outcome

We investigated whether the intervention effects were sensitive to controlling for the type of the device that the experiment was completed on, as well as whether it was sensitive to controlling for participants who watched three potentially harmful in a row. We ran an additional sensitivity analysis without observations of videos that were not watched by participants. These videos were not watched, because the videos did not auto-play and participants skipped these videos without playing them. Further information around this issue is provided in section 7.2.2.

Secondary analyses

Secondary outcome 2.1 (in Table 2) was analysed in the same way as the primary outcome (1.1 in Table 2), to determine whether the effect estimates of the interventions were sensitive to how the skipping behaviour was measured.

The intention was to analyse secondary outcome 2.2 (in Table 2) using the same model as the one described in the primary analysis. However, by comparing the expected distribution of zeros (derived by running 1,000 simulations of scaled residuals using a fitted mixed-effects logistic regression model) against the observed values (from the collected data), we found more zeros in the data than expected (zero-inflation). Thus, the secondary outcome 2.2 was analysed in the same way as the primary outcome in the Ofcom 1 experiment: using a zero-inflated Poisson regression model, with a sensitivity check using a Hurdle model.

Secondary outcome 2.3 (in Table 2) was analysed using a similar model to the primary outcome (1.1 in Table 2). The difference in the modelling approach was that for secondary outcome 2.3 the model was a linear mixed-effects model (because the outcome was continuous), rather than a logistic mixed-effects model that was used for the binary outcome variable that constituted the primary outcome in

this study. In addition, the model was fitted with robust standard errors (because the assumption of homogeneity of variance was violated).

Similarly, we wanted to analyse secondary outcome 2.4 (in Table 2) using mixed-effects models. However, due to random effects variances being estimated as 0 (singularity), we simplified the model. We analysed this outcome using a linear regression with robust standard errors.

Three of the attitudinal questions (Secondary outcome 2.5 (in Table 2)) were analysed using inferential tests, and ordinal models, to compare the responses of the participants in the intervention arms compared to those in the control arm. The set of questions concerning whether participants regretted watching potentially harmful videos, were analysed using a mixed-effects ordinal regression model.

Descriptive statistics

We also conducted exploratory analysis relating to how the primary outcome measure (1.1 in Table 2) or other secondary outcomes may vary across multiple sub-level groupings like age groups or self-reported gender to provide additional insights into participant behaviour. However, because this research was not designed to explore potential demographic differences across participants, it was unknown whether this study was sufficiently powered to detect any potential effects. As such, any conclusions drawn from this reporting should not be interpreted as representative of the population of Profiles panel members who were VSP users.

6.2. Statistical power

To run power simulations for logistic mixed-effects models, assumptions about the variance and standard deviation parameters of the random effects were required. To obtain meaningful estimates of power using power simulations, these assumptions should be grounded in prior research. Thus, the estimates of the parameters of variation in the probability of skipping between participants (random intercept for participants (σ_1)) and in the probability of skipping between potentially harmful videos (random intercept for potentially harmful videos (σ_2)) were based on the estimates found in the Ofcom 2 trial. Effect sizes of 8%, 9% and 10% in the probability of skipping - per intervention arm compared to the control arm - were assumed under different scenarios. As with the previous behavioural experiments for Ofcom, the minimum detectable effect size of interest was set at 10%.

Table 3 shows the estimates of power under different model assumptions, given 1,000 simulations per scenario. It was unlikely that the estimates under any scenario would be the same as the ones obtained using models given the collected data. Thus, the estimates of power to detect the effect of the intervention, given the scenarios considered, were not an exact representation of the true effect of the interventions (in the context of our study). However, they were likely to be reasonably close since we were using the same sampling strategy and potentially harmful videos, as in Ofcom 2. In the case of a null effect, the estimates below may provide evidence as to where the null effect might have come from (for example, small effect of the intervention or large variation due to individual differences).

Following the results of the power simulations, a decision was taken to recruit a sample of 2,800 participants, resulting in 700 participants per treatment arm. This was informed by Scenario 12 in Table 3. Scenario 12 shows that, given our assumptions, 2,800 participants were needed to ensure that the trial was sufficiently powered to detect a minimum effect size of 10%. Specifically, under Scenario 12, the power to reject the null hypothesis of there being no difference between arms was 88% (which is higher the conventional threshold, used in power simulations, of 80% at $\propto = 0.05$).

Table 3. Power to detect an effect of specified size, by scenario

Scenario	Sample Size (3 videos each)	Effect	σ_1	σ_2	σ_3[16]	Power
1	1600	8%	1.94	0.46	0.1	40%
2	1600	9%	1.94	0.46	0.1	52%
3	1600	10%	1.94	0.46	0.1	60%
4	2000	8%	1.94	0.46	0.1	49%
5	2000	9%	1.94	0.46	0.1	63%
6	2000	10%	1.94	0.46	0.1	73%

[16] The parameter was added in power simulations to be conservative (which is useful in the context where we can expect the interface not to work equally well for every participant), but it is not something that can be estimated by the model. Observation level variability parameter is not required in logistic models - the probability parameter captures this from observational variability in the draw from the binomial distribution (the probability parameter captures both location of the expectation of the mean and how much variance there will be). Thus, σ_3 is adding noise that cannot be modelled by the logistic model. It reflects variance that is above and beyond observation level noise than would be expected by the logistic model.

7	2400	8%	1.94	0.46	0.1	55%
8	2400	9%	1.94	0.46	0.1	71%
9	2400	10%	1.94	0.46	0.1	79%
10	2800	8%	1.94	0.46	0.1	66%
11	2800	9%	1.94	0.46	0.1	78%
12	2800	10%	1.94	0.46	0.1	88%

7. Results

7.1. Randomisation and balance between arms

The randomisation process resulted in relatively balanced split of participants according to demographic variables within each treatment arm. For example, the median age of participants across arms ranged from 40 to 41 (Table 4).[17]

Table 4. Split of participants by age, gender, and socio-economic group (SEG), variables across trial arms

	Age (Median)	Gender (% Male)	SEG (% ABC1)
Arm 1 Control	40	48.1	56.6
Arm 2 Active choice	40	51.1	55.2
Arm 3 Auto-skip	41	47.3	56.3
Arm 4 Auto-play	41	48.6	55.9

Note: ABC1 refers to upper middle class (A), middle class (B), and lower middle class (C1)

The distribution of device operating system used to complete the experiment was relatively balanced across each treatment arm (Table 5). Slightly fewer participants completed the experiment on Android devices (and slightly more on iOS devices) in the control arm, compared to the treatment arms. (Refer to Appendix B, in section 10, for the demographic breakdown of participants by device type.)

Table 5. Split of participants by device operating system, by arm

Device operating system	Arm 1 Control (%)	Arm 2 Active choice (%)	Arm 3 Auto-skip (%)	Arm 4 Auto-play (%)
Android	36.4%	39.1%	40.4%	39.4%
iOS	29.4%	27.2%	25.3%	25.4%
Windows	24%	26.8%	23.6%	26.1%
macOS	5.6%	4.6%	6.0%	5.3%
iPadOS	1.7%	0.6%	1.1%	0.7%
ChromeOS	2.3%	1.4%	2.6%	2.3%
Linux	0.4%	0.3%	1.0%	0.1%
Unknown	0.1%	0%	0%	0.6%

7.2. Primary Outcome Analysis

7.2.1. Skipping of potentially harmful videos

The mean observed probability of skipping potentially harmful videos was 65% in the control arm, 67% in Arm 2 – Active choice, 78% in Arm 3 – Auto-skip, and 65% in Arm 4 – Auto-play.[18] Table 6 shows that the odds of skipping the potentially harmful videos were significantly higher in Arm 3 – Auto-skip compared to the control arm. The whole model, including random effects, accounted for 52% of the variance in the probability of skipping potentially harmful videos (pseudo delta R^2 was 0.52).[19]

Table 6. Model-based estimates of the odds of skipping of potentially harmful videos

	Odds Ratios	95% CI	*z*-value	*P*
Intercept	3.13	2.37 – 4.13	8.085	< 0.001
Arm 2 Active choice	1.12	0.83 – 1.52	0.755	0.450
Arm 3 Auto-skip	3.25	2.37 – 4.47	7.292	< 0.001

[17] Note that perfect balance does not have to be achieved for mixed-effects models to work well. This is because groups with less data will automatically be shrunk towards overall mean values. Consequently, mixed-effects models are well suited to unbalanced designs.
[18] Mean observed probabilities were calculated using observed values (from the collected data). These are reported as they are typically easier to understand, however they do not directly relate to model estimates and cannot be calculated using estimated, reported, Odds Ratios.
[19] Note that random effects models consider participant and video variability, thus the reported estimates are not driven by one particular video or by one particular participant.

Arm 4 Auto-play	0.95	0.70 – 1.28	-0.334	0.738

Note: Arm 1 – Control is the reference level other arms are compared against

Adjusting for multiple comparisons, the odds of skipping potentially harmful videos were higher in Arm 3 – Auto-skip compared to any other arm (see Table 7).

There were slight differences in the proportion of device types used to complete the experiment between the control and intervention arms (see Table 5). Consequently, we re-ran the primary model with a device type covariate to control for any differences due to differences in proportions, between control and intervention arms, of device types used to complete the experiment. A model with all device types did not converge.[20] For this model to converge, 54 observations from uncommon devices / operating systems were removed from the analysis,[21] and observations relating to iOS, MacOS, and iPadOS categories were grouped together.[22] The reported effects were not sensitive to controlling for the device type.

In addition, the reported effects were not sensitive to the inclusion of variable indicating whether participants watched three videos in a row or not; the probability of skipping was not significantly different for participants who were shown three potentially harmful videos in a row compared to those who did not.

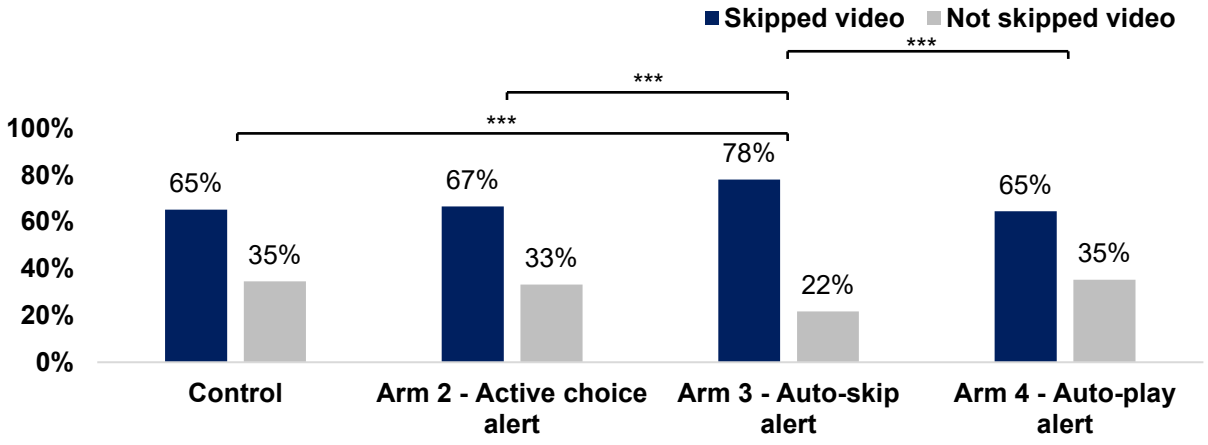
Table 7. Estimates of the odds of skipping of potentially harmful videos (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*-value	*P*
Arm 1 Control vs Arm 2 Active choice	1.12	0.75 – 1.67	0.755	1*
Arm 1 Control vs Arm 3 Auto-skip	3.25	2.14 – 4.93	7.292	< 0.001
Arm 1 Control vs Arm 4 Auto-play	0.95	0.64 – 1.41	-0.334	1*
Arm 2 Active choice vs 3 Auto-skip	2.90	1.91 – 4.39	6.581	< 0.001
Arm 2 Active choice vs 4 Auto-play	0.85	0.57 – 1.26	-1.089	1*
Arm 3 Auto-skip vs 4 Auto-play	0.29	0.19 – 0.44	-7.604	< 0.001

* Approximately (rounding error)

Figure 11 shows the percentage of participants skipping potentially harmful videos, by arm. Significant differences, as estimated using the primary outcome model, are shown using horizontal lines.

Figure 11. Percentage of skipped and not skipped potentially harmful videos, by arm (multiple comparisons adjusted; *** p < 0.001; not sensitive to controlling for device type; modelled using a mixed-effects logistic model)



[20] The implications of a model not converging are that the estimates of such a model cannot be trusted. This is because such a model's estimates are likely to be inaccurate (not precise), unreliable (not consistent), and/or biased (distorted).

[21] By rarely used devices, we mean 'Unknown' (constituting 15 observations, therefore 5 participants) and 'Linux' (constituting 39 observations, therefore 13 participants) devices. The observations relating to these devices were removed, because the inclusion of them led to lack of model convergence. This is because effect estimates for some arms could not have been reliably estimated. For example, Arm 4 did not contain any participants who completed the experiment on an 'Unknown' device. Similarly, only 1 participant in Arm 4 completed the experiment using a 'Linux' device.

[22] Observations relating to iOS, MacOS, and iPadOS were grouped together for three reasons. First, only 29 participants completed the experiment using iPadOS, and the distribution was not even between arms (for example, only 4 participants completed the experiment using iPadOS in Arm 4). Second, the issue of videos not auto-playing for participants seemed to affect all Apple devices to a similar extent (see section 7.2.2). Consequently, it made sense to treat them the same. Third, the model did not converge without grouping these three device types together.

7.2.2. Sensitivity check

Whilst user testing a separate online trial, it became apparent that for a minority of participants in certain circumstances, videos would not begin playing automatically[23]. This issue did not affect the alert message stage of the trial but may have impacted user experience of the video interface. Thus, we conducted sensitivity analyses to determine whether the reported effects were affected by this issue. Specifically, we excluded participant observations in which a potentially harmful video was skipped via the interface (after the alert message) with a watch time of 0 seconds[24]. In total, this led to the exclusion of 341 observations. We re-ran the primary model without these 341 observations to examine intervention effects under conditions where all participants see videos.

In addition, we also re-ran the primary model with a device type covariate to control for any differences due to differences in proportions, between control and intervention arms, of device types / operating systems used to complete the experiment. For this model to converge, 50 observations from uncommon devices / operating systems were removed from the analysis (see footnotes 19 and 20).[25]

The observed probability of skipping potentially harmful videos in this adjusted primary model was 61% in the control arm, 66% in Arm 2 – Active choice, 78% in Arm 3 – Auto-skip, and 64% in Arm 4 – Auto-play. Table 8 shows that the odds of skipping the potentially harmful videos were significantly higher in Arm 2 – Active choice compared to the control arm, and in Arm 3 – Auto-skip compared to the control arm.

Table 8. Sensitivity check model-based estimates of the odds of skipping of potentially harmful videos

	Odds Ratios	95% CI	*z*-value	*P*
Intercept	2.27	1.67 – 3.10	5.216	< 0.001
Arm 2 Active choice	1.53	1.12 – 2.10	2.678	0.007
Arm 3 Auto-skip	4.58	3.29 – 6.38	9.008	< 0.001
Arm 4 Auto-play	1.24	0.91 – 1.70	1.374	0.170

Note: Arm 1 – Control is the reference level other arms are compared against

Adjusting for multiple comparisons, the odds of skipping potentially harmful videos were higher in Arm 3 – Auto-skip compared to any other arm, and higher in Arm 2 – Active choice compared to the control arm (see Table 9).

The sensitivity check appears to indicate that the probability of skipping in Arm 2 – Active Choice could be significantly different from Arm 1 – Control but it is not possible to say definitively that this is due to correcting for the device type issue because of the need to remove the 50 observations relating to rarely used devices as well.

Table 9. Estimates of the odds of skipping of potentially harmful videos (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*-value	*P*
Arm 1 Control vs Arm 2 Active choice	1.53	1.02 – 2.31	2.678	0.044
Arm 1 Control vs Arm 3 Auto-skip	4.58	2.97 – 7.07	9.008	< 0.001
Arm 1 Control vs Arm 4 Auto-play	1.24	0.83 – 1.87	1.374	1*
Arm 2 Active choice vs 3 Auto-skip	2.99	1.96 – 4.56	6.634	< 0.001
Arm 2 Active choice vs 4 Auto-play	0.81	0.54 – 1.22	-1.326	1*
Arm 3 Auto-skip vs	0.27	0.18 – 0.42	-7.862	< 0.001

[23] The issue mainly affected participants who accessed the trial on a device running iOS (the operating system for iPhones, n = 182 observations). However, other Apple devices, running macOS and iPadOS, were also affected. We believe that this is a design choice by Apple, and that this cannot be overridden in this online environment.

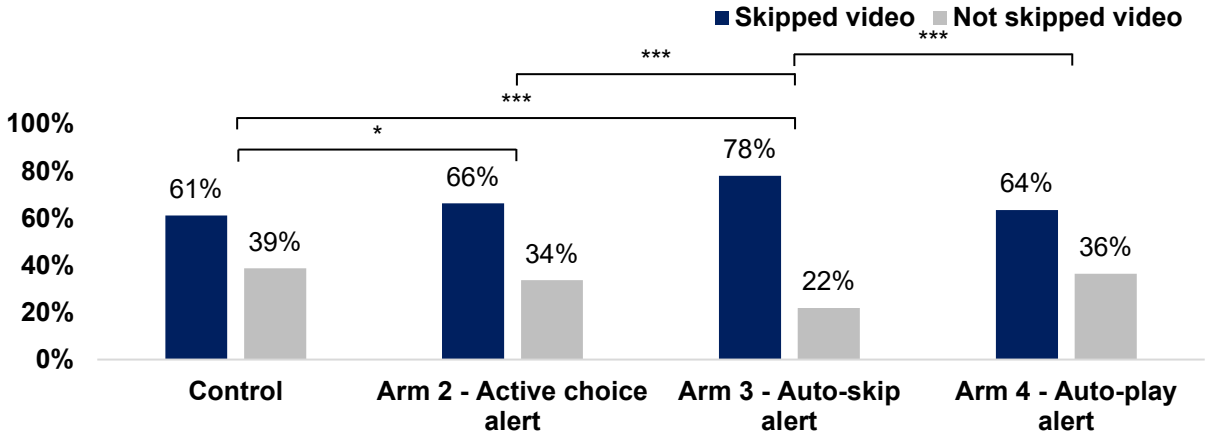
[24] Device type or operating system was not included as criteria for the exclusion of observations, as this was also an issue for a minority of participants using other devices / operating systems, such as Android (n = 44 observations), ChromeOS (n = 2 observations), iPadOS (n = 8 observations), macOS (n = 33 observations), unknown device type (n = 4 observations), and Windows (n = 68 observations).

[25] The number of observations removed from this dataset was 50, as opposed to the 54 observations that were removed from the dataset used in the analysis reported in section 7.2.1. This is because the dataset used for sensitivity analysis did not have the 341 observations in which a potentially harmful video was skipped via the interface with a watch time of 0 seconds. Out of these 341 observations, 4 were related to participants who completed the experiment using either 'Unknown' or 'Linux' devices.

* Approximately (rounding error)

Figure 12 shows the percentage of participants skipping potentially harmful videos, by arm. Significant differences, as estimated using a model without 341 observations, are shown using horizontal lines.

Figure 12. Percentage of skipped and not skipped potentially harmful videos, by arm – sensitivity with some observations removed (multiple comparisons adjusted; * p < 0.05; *** p < 0.001; the difference between Arm 1 – Control and Arm 2 – Active choice was sensitive to controlling for device type; modelled using a mixed-effects logistic model)



7.3. Secondary Outcome Behavioural Analysis

7.3.1. Skipping of neutral videos

As part of our sensitivity analysis, the same model as in section 7.2 was run on the skipping data of neutral videos. It showed no differences in the probability of skipping neutral content across all four arms of the trial.

The observed probability of skipping videos was 60% in the control arm, 57% in Arm 2 – Active choice, 58% in Arm 3 – Auto-skip and in Arm 4 – Auto-play. There were no significant differences in the odds of skipping neutral videos between the intervention arms and the control arm (Table 10). (These findings were not sensitive to the exclusion of 341 observations in which a potentially harmful video was skipped via the interface with a watch time of 0 seconds.) The whole model, including random effects, accounted for 51% of the variance in the probability of skipping neutral videos.

Table 10. Model-based estimates of the odds of skipping of neutral videos

	Odds Ratios	95% CI	*z*-value	*P*
Intercept	1.95	0.86 – 4.41	1.593	0.111
Arm 2 Active choice	0.84	0.65 – 1.09	-1.319	0.187
Arm 3 Auto-skip	0.90	0.69 – 1.17	-0.776	0.438
Arm 4 Auto-play	0.91	0.70 – 1.19	-0.678	0.498

Note: Arm 1 – Control is the reference level other arms are compared against

Adjusting for multiple comparisons, there were no significant differences in the odds of skipping neutral videos between any arm (see Table 11). Thus, alert messages had no impact on the skipping probability of skipping of neutral content. Figure 13 visualises this by showing the percentage of participants skipping neutral videos, by arm.

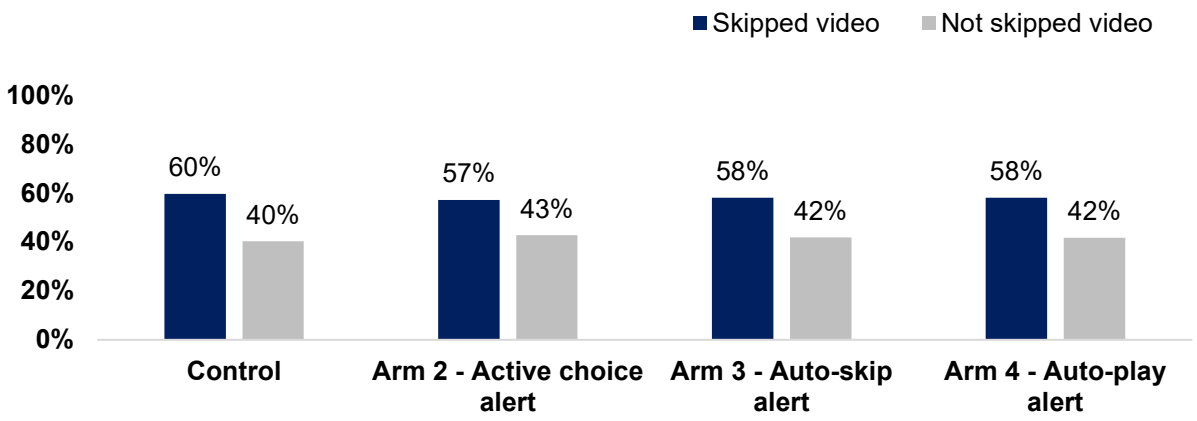
Table 11. Estimates of the odds of skipping of neutral videos (p values and CIs corrected for multiple comparisons using the Bonferroni correction)

Comparison	Odds Ratios	95% CI	*z*-value	*P*
Arm 1 Control vs Arm 2 Active choice	0.84	0.60 – 1.18	-1.319	1*
Arm 1 Control vs Arm 3 Auto-skip	0.90	0.64 – 1.27	-0.776	1*
Arm 1 Control vs Arm 4 Auto-play	0.91	0.65 – 1.29	-0.678	1*

Arm 2 Active choice vs 3 Auto-skip	1.07	0.76 – 1.51	0.542	1*
Arm 2 Active choice vs 4 Auto-play	1.09	0.77 – 1.54	0.638	1*
Arm 3 Auto-skip vs 4 Auto-play	1.01	0.72 – 1.43	0.097	1*

* Approximately (rounding error)

Figure 13. Percentage of skipped and not skipped neutral videos, by arm (multiple comparisons adjusted; modelled using a mixed-effects logistic model)

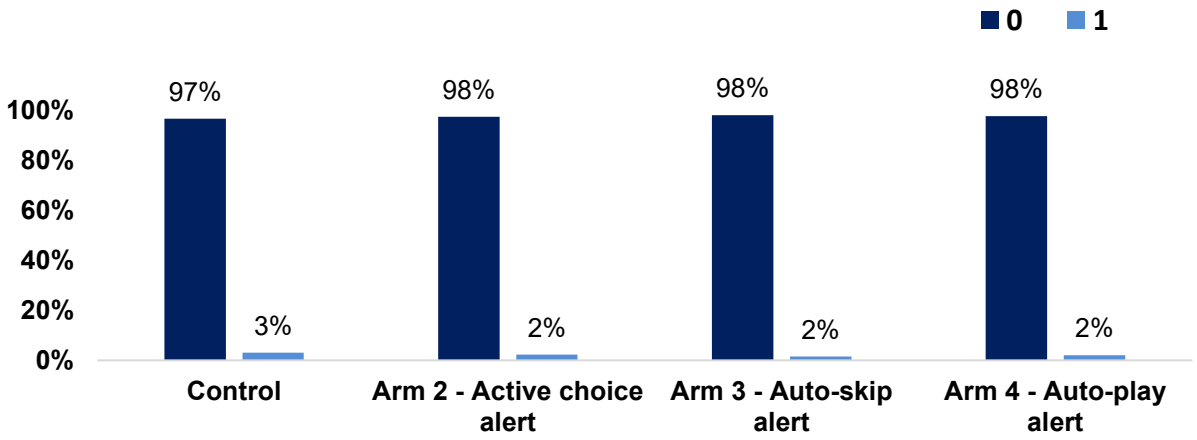


7.3.2. Reporting of potentially harmful videos

We analysed the reporting of potentially harmful videos using a mixed-effects logistic regression (with random intercept for participants, only), zero-inflated Poisson regression model, and a Hurdle model. Additional models with zero-inflation components were fitted because the observed number of zeros exceeded the simulated distribution of zeros. Specifically, the ratio of observed to simulated zeros was 1.14, a significant deviation from 1 ($p < 0.001$). (A ratio greater than 1 can indicate zero inflation.)

Observed mean probability of reporting was 3% in the control arm and 2% in the intervention arms. (The counts of reports were low, with 65 potentially harmful videos reported in the control arm, 48 in Arm 2 – Active choice, 34 in Arm 3 – Auto-skip, and 41 in Arm 4 – Auto-play.) Adjusting for multiple comparisons, we did not find any differences in the probability of reporting between arms, using any model. Figure 14 shows the percentage participants reporting potentially harmful content videos, by arm.

Figure 14. Percentage of reported (1) and not reported (0) potentially harmful videos, by arm



7.3.3. View times of potentially harmful videos

Participants in Arm 1 – Control spent less time (in seconds) watching potentially harmful videos (M = 23.72; Mdn = 18.10) than participants in any of the intervention arms (excluding observations of videos skipped at the alert stage). See Table 12. The fact that participants in the intervention arms who had chosen to watch the potential harmful videos, watched them for longer than participants in the control arm suggests that they were more comfortable / tolerant of that content while those participants who were less comfortable / tolerant had opted to skip at the alert message stage. In the control group, participants who were less comfortable / tolerant with the potentially harmful content would not have received any warning about the nature of the content and so would only be able to

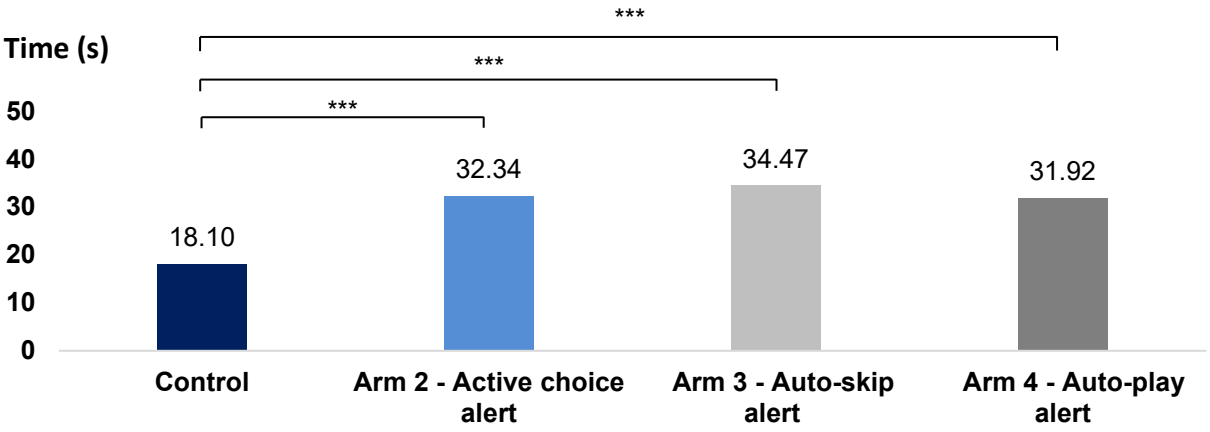
respond by skipping content after they had started to watch it. The result of this would be to reduce the average length of time spent watching the potentially harmful content in the control group.

Table 12. Mean and median view times (in seconds) of potentially harmful videos, by arm

Arm	Mean	SD	Median
Arm 1 Control	23.72	20.23	18.10
Arm 2 Active choice	29.22	18.94	32.34
Arm 3 Auto-skip	29.53	18.75	34.47
Arm 4 Auto-play	28.63	19.34	31.92

Figure 15 shows the median time (in seconds) participants spent watching each potentially harmful video, in different arms (excluding participants who skipped the potentially harmful video at the pop-up alert stage).

Figure 15. Median view times (in seconds) of potentially harmful videos, by arm



The time spent watching potentially harmful videos was analysed using a robust mixed-effects linear regression model. Excluding participants who skipped the videos at the alert stage, the length of viewing time across potentially harmful videos was predicted to be higher in the intervention arms compared to the control arm (Table 13).

Table 13. Model-based estimates of view times (in seconds) of potentially harmful videos

	Estimate	SE	*t*-value	*p*
Intercept	23.67	2.68	8.826	< 0.001
Arm 2 Active choice	5.60	1.08	5.197	< 0.001
Arm 3 Auto-skip	5.70	1.16	4.929	< 0.001
Arm 4 Auto-play	4.81	1.05	4.567	< 0.001

Note: Arm 1 – Control is the reference level other arms are compared against

The differences were not sensitive to correcting for multiple comparisons (see Table 14). In addition, the differences were not sensitive to re-running the analysis without the 341 observations of potentially harmful videos that were skipped by participants without being played. (Although the magnitude of the differences was smaller.)

Table 14. Estimates of view times (in seconds) of potentially harmful videos (p values corrected for multiple comparisons using the Bonferroni correction)

Comparison	Estimate	SE	*t*-value	*p*
Arm 1 Control vs Arm 2 Active choice	5.60	1.08	5.197	< 0.001
Arm 1 Control vs Arm 3 Auto-skip	5.70	1.16	4.929	< 0.001
Arm 1 Control vs Arm 4 Auto-play	4.81	1.05	4.567	< 0.001
Arm 2 Active choice vs 3 Auto-skip	0.10	1.20	0.081	1*
Arm 2 Active choice vs 4 Auto-play	-0.79	1.19	-0.721	1*
Arm 3 Auto-skip vs 4 Auto-play	-0.89	1.17	-0.756	1*

*Approximately (rounding error)

7.3.4. Response times (in seconds) to alert messages

Participants in Arm – 2 Active choice had higher mean (4.84) and median (3.15) response times to alert messages (by either clicking "proceed" or "skip") than participants in Arm 3 – Auto-skip (M = 3.16; Mdn = 2.96) and Arm 4 – Auto-play (M = 3.13; Mdn = 2.90). See Table 15 for mean and median interaction times in all arms. (Note that we excluded observations of auto-skips and auto-plays, as well as those of 3 participants who took more than 1,000 seconds to interact with the alert message.)[26]

Table 15. Mean and median times (in seconds) to interact with the alert message, per intervention arm

Arm	Mean	SD	Median
Arm 2 Active choice	4.84	13.46	3.15
Arm 3 Auto-skip	3.16	1.91	2.96
Arm 4 Auto-play	3.14	2.15	2.90

The response times to alert messages was analysed using a robust linear regression model. The length of time to interact with the alert message was estimated to be lower in the Arms 3 and 4, compared to Arm 2 – Active choice (Tables 16 and 17).

Table 16. Model-based estimates of response times (in seconds) to alert messages

	Estimate	SE	*t*-value	*p*
Intercept	3.30	0.03	107.883	< 0.001
Arm 3 Auto-skip	-0.23	0.05	-4.959	< 0.001
Arm 4 Auto-play	-0.28	0.05	-5.498	< 0.001

Note: Arm 2 – Active choice is the reference level other arms are compared against

The differences were not sensitive to correcting for multiple comparisons, see Table 17. (The significance of the differences was also not sensitive to the choice of the model.) In addition, the differences were not sensitive to re-running the analysis without the 341 observations of potentially harmful videos that were skipped by participants without being played. (Although the magnitude of the differences was smaller.)

Table 17. Estimates of response times (in seconds) to alert messages (p values corrected for multiple comparisons using the Bonferroni correction)

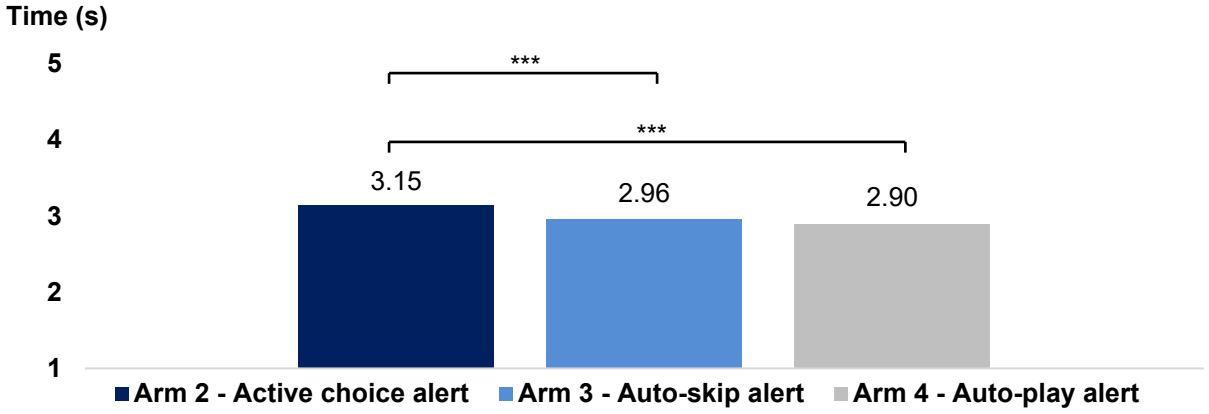
Comparison	Estimate	SE	*t*-value	*p*
Arm 2 Active choice vs 3 Auto-skip	-0.23	0.05	-4.959	< 0.001
Arm 2 Active choice vs 4 Auto-play	-0.28	0.05	-5.498	< 0.001
Arm 3 Auto-skip vs 4 Auto-play	-0.05	0.05	0.851	1*

*Approximately (rounding error)

Figure 16 shows the median times (in seconds) it took participants to interact with each alert message, per intervention arm (Arm 2 – Active choice = 3.15; Arm 3 – Auto-skip = 2.96; Arm 4 – Auto-play = 2.90).

[26] Times of auto-skips and auto-plays were excluded because these were not response times of participants. The decision to use the 1,000 seconds threshold to exclude participants was largely arbitrary. We did not want to exclude any observations, because we do not know whether they are recording genuine behaviour or are errors of measurement. 1,000 seconds threshold was used, because the next longest observation was 296 seconds. Thus, we believed that the three observations of over 1,000 seconds were qualitatively different than the rest. The inference is not sensitive to a lower threshold of excluding all observations that were 100 seconds in length or longer (note that the median times are the same given a threshold of 1,000 and a threshold of 100).

Figure 16. Median times (in seconds) spent at the alert message stage, per intervention arm



7.4. Secondary Outcome Attitudinal Analysis

7.5.1. Perceptions of alert messages[27]

We found no significant differences in self-reported perceived usefulness, annoyingness, and distraction between participants exposed to different alert messages. In addition, we did not find any significant differences in self-reported levels of regret after choosing to watch potentially harmful videos, between participants exposed to different types of alert messages.

The intervention arms had very similar mean usefulness ratings (Arm 2 – Active choice = 3.84; Arm 3 – Auto-skip = 3.79; Arm 4 – Auto-play = 3.83) and the same median usefulness rating of 4. See Table 18. A Kruskal-Wallis test, and an ordinal regression, were conducted to identify if there were significant differences between treatment arms in their perceptions of usefulness of the warning. No significant differences between intervention arms were found.

Table 18. Mean and median values of usefulness, per intervention arm

Arm	Mean	SD	Median
Arm 2 Active choice	3.84	1.13	4
Arm 3 Auto-skip	3.79	1.08	4
Arm 4 Auto-play	3.83	1.12	4

The intervention arms had very similar mean annoyingness ratings (Arm 2 – Active choice = 2.41; Arm 3 – Auto-skip = 2.56; Arm 4 – Auto-play = 2.45) and the same median annoyingness rating of 2. See Table 19. Adjusting for multiple comparisons, no significant difference in perceptions of the 'annoyingness' of warnings was also detected.

Table 19. Mean and median values of annoyingness, per intervention arm

Arm	Mean	SD	Median
Arm 2 Active choice	2.41	1.21	2
Arm 3 Auto-skip	2.56	1.24	2
Arm 4 Auto-play	2.45	1.20	2

The intervention arms had very similar mean distraction ratings (Arm 2 – Active choice = 2.38; Arm 3 – Auto-skip = 2.50; Arm 4 – Auto-play = 2.48) and the same median distraction rating of 2. See Table 20. No significant difference in perceptions of the 'distraction' of warnings was detected.

Table 20. Mean and median values of distraction, per intervention arm

Arm	Mean	SD	Median
Arm 2 Active choice	2.38	1.12	2
Arm 3 Auto-skip	2.50	1.16	2
Arm 4 Auto-play	2.48	1.16	2

[27] Excludes 'Don't know' responses.

Those who were warned about potentially harmful video content but chose to watch the video anyway were asked the extent to which they regretted their choice. On a scale of 1 ('Disagree strongly') to 5 ('Agree strongly'). Table 21 shows that the intervention arms had very similar mean regret ratings (Arm 2 – Active choice = 2.76; Arm 3 – Auto-skip = 2.77; Arm 4 – Auto-play = 2.83) and the same median regret rating of 3.

Table 21. Mean and median values of regret, per intervention arm

Arm	Mean	SD	Median
Arm 2 Active choice	2.76	1.30	3
Arm 3 Auto-skip	2.77	1.29	3
Arm 4 Auto-play	2.83	1.30	3

To analyse this outcome measure, a variation of a mixed-effects ordinal regression model, cumulative link mixed-effects model (CLMM), was used.[28] Using CLMM, there were no significant differences in the odds of regretting viewing over all potentially harmful videos between the intervention arms (Table 22).

Table 22. Model-based estimates of the odds of regretting watching potentially harmful videos

Arm	Odds Ratios	95% CI	*z*-value	*P*
Arm 3 Auto-skip	1.12	0.83 – 1.52	0.740	0.459
Arm 4 Auto-play	1.27	0.96 – 1.67	1.671	0.095

Note: The Odds Ratios for Arms 3 and 4 are evaluated against Arm 2 – Active choice

7.7. Descriptive statistics

We also conducted exploratory analysis relating to how the primary outcome measure (1.1 in Table 2) or other secondary outcomes may vary across multiple sub-level groupings like age groups or self-reported gender to provide additional insights into participant behaviour. However, this research was not designed to explore potential demographic differences across participants. As a result, it is not known whether this study was sufficiently powered to detect any potential effects. As such, any conclusions drawn from this reporting should not be interpreted as representative of the population of panel members who are VSP users.

7.7.1. Skipping of potentially harmful videos by age and gender

Table 23 shows the observed probability of skipping potentially harmful videos by age. Participants in the age group of 18-24 had lower mean observed probability of skipping potentially harmful videos (0.64) than participants of any other age group. Participants in the age group of 55-69 had a higher mean observed probability of skipping potentially harmful videos (0.72) than any other age groups.[29]

Table 23. Mean observed probability of skipping potentially harmful videos, by age group

Age group	Mean	SD
18-24	0.64	0.48
25-39	0.69	0.47
40-54	0.69	0.46
55-69	0.72	0.45

Table 24 shows the observed probability of skipping of potentially harmful videos by gender. Female participants had a higher mean observed probability of skipping potentially harmful videos (0.71) than male participants (0.66).

[28] Using ordinal models to model ordinal data enables more accurate estimation of the effects than any model which assumes metric or categorical responses. A cumulative model assumes that the observed ordinal outcome variable represents the categorization of a latent continuous variable. It models this categorization by assuming that there are several thresholds at which the outcome variable is partitioned. This categorization is commonly used to model Likert-scale data, when ordered labels are used to collect judgements about a potentially continuous latent variable. CLMM, a cumulative model with fixed and random effects, was appropriate for two reasons. First, there was a crossed random effects structure: each participant was asked to rate their regret for every potentially harmful video they chose to watch. Second, the outcome variable (regret) can be considered an ordinal scale whereby there is ordering of the levels (from 1 to 5) and an upper (5) and lower (1) limit for each writing task (CLMMs allow to account for the potential ceiling and floor effects imposed by these limits, in a way that standard analyses do not).

[29] These age groups have come from the quotas that were used for the recruitment of participants (section 2.1).

Table 24. Mean observed probability of skipping potentially harmful videos, by gender

Gender	Mean	SD
Male	0.66	0.47
Female	0.71	0.45
Other	0.58	0.50
Prefer not to say	1	0

7.7.2. Probability of reporting potentially harmful videos by age or gender

Table 25 shows the observed probability of reporting potentially harmful videos by age. (Note that potentially harmful videos were reported only 188 times. Thus, no conclusions should be drawn based on the reported descriptive statistics.) Participants in the age group of 18-24 had lower mean observed probability of reporting potentially harmful videos (0.01) than participants of any other age group. However, due to the low count of reports, and low mean observed probabilities of reporting between age groups, these estimates should be interpreted with caution.

Table 25. Mean observed probability of reporting potentially harmful videos, by age group

Age group	Mean	SD
18-24	0.01	0.12
25-39	0.03	0.16
40-54	0.02	0.14
55-69	0.03	0.16

Table 26 shows the observed probability of reporting potentially harmful videos by gender. There were no noticeable differences in the observed probability of reporting potentially harmful videos between male (M = 0.02) and female (M = 0.02) participants. Only 33 observations (out of 8,403) are attributable to participants who identified as “Other”. Thus, no comparisons can be reliably made using this group of participants.

Table 26. Mean observed probability of reporting potentially harmful videos, by gender

Gender	Mean	SD
Male	0.02	0.14
Female	0.02	0.15
Other	0.18	0.39
Prefer not to say	0	0

7.7.3. View time of potentially harmful videos

Table 27 shows the mean and median times (in seconds) participants spent watching potentially harmful videos, by age. Participants in the age group of 18-24 spent more time watching potentially harmful videos than participants of any other age group (M = 28.32; Mdn = 38.97). Median times were the lowest amongst the age groups of 40-54 (19.15) and 55-69 (18.62).

Table 27. Mean and median view times (in seconds) of potentially harmful videos, by age group

Age group	Mean	SD	Median
18-24	28.32	20.71	38.97
25-39	27.02	20.15	28.03
40-54	27.28	19.15	19.15
55-69	27.36	18.62	18.62

Table 28 shows the mean and median times (in seconds) participants spent watching potentially harmful videos, by gender. Male participants had longer watch times (M = 28.71; Mdn = 33.98) of potentially harmful videos than female participants (M = 26.00; Mdn = 23.86).

Table 28. Mean and median view times (in seconds) of potentially harmful videos, by gender

Gender	Mean	SD	Median
Male	28.71	19.46	33.98
Female	26.00	19.63	23.86
Other	33.30	19.54	41.68
Prefer not to say	7.21	8.75	4.33

7.7.4. Response times (in seconds) to alert messages

Table 29 shows the mean and median times (in seconds) it took participants to interact with the alert message, by age. (Note that we excluded observations of videos that auto-skipped or auto-played.) There were no reliable differences in the time it took participants to interact with the alert messages between age groups, when considering mean (from 3.37 for 18-24-year-olds to 4.13 for 40-54-year-olds) and median (from 2.84 for 25-39-year-olds to 3.30 for 55-69-year-olds) times, together.

Table 29. Mean and median response times (in seconds) to alert messages, by age group

Age group	Mean	SD	Median
18-24	3.37	3.18	2.95
25-39	4.04	10.67	2.84
40-54	4.13	11.37	3.04
55-69	3.67	3.58	3.30

Table 30 shows the mean and median times (in seconds) it took participants to interact with the alert message, by gender. (We excluded observations of videos that auto-skipped or auto-played.) Male participants (M = 3.83; Mdn = 2.99) were slightly faster to interact with the alert messages than female participants (M = 3.95; Mdn = 3.07), but this difference appears to be negligible.

Table 30. Mean and median response times (in seconds) to alert messages, by gender

Gender	Mean	SD	Median
Male	3.83	9.21	2.99
Female	3.95	8.93	3.07
Other	3.16	1.81	2.41
Prefer not to say	3.37	1.04	3.26

7.7.5. Engagements and view times

After excluding observations in which participants skipped potentially harmful videos at the alert stage, we observed minor differences between participants in the control and intervention conditions (Table 31). Specifically, participants in the intervention arms had slightly higher mean observed probability of commenting (Arm 2 – Active choice = 0.05; Arm 3 – Auto-skip = 0.06; Arm 4 – Auto-play = 0.04) than participants in the control arm (0.03). This pattern was repeated in disliking, where participants in the intervention arms had higher mean observed probabilities (Arm 2 – Active choice = 0.27; Arm 3 – Auto-skip = 0.27; Arm 4 – Auto-play = 0.25) than participants in the control arm (0.22). Thus, all alert messages increased the mean observed probabilities of commenting and disliking potentially harmful videos that were watched (including partially watched). Conversely, there were no differences in mean observed probabilities of potentially harmful videos that were ‘Liked’ (0.09 in all arms) and ‘Shared’ (0.01 in all arms).

Table 31. Engagements (mean observed probabilities) with potentially harmful videos, by arm (Excluding observations of skipped videos at the alert stage)

Arm	Liked	Shared	Commented	Disliked
Arm 1 Control	0.09	0.01	0.03	0.22

Arm 2 Active choice	0.09	0.01	0.05	0.27
Arm 3 Auto-skip	0.09	0.01	0.06	0.27
Arm 4 Auto-play	0.09	0.01	0.04	0.25

Excluding observations of skipped videos at the alert stage, the mean total view time of potentially harmful videos was longer for participants in the intervention arms (Arm 2 – Active choice = 29.22; Arm 3 – Auto-skip = 29.53; Arm 4 – Auto-play = 28.63) than for participants in the control arm (23.72). (See Table 32.) This pattern of mean total view times was repeated for neutral videos, but the magnitude of differences was smaller (Arm 1 – Control = 24.33; Arm 2 – Active choice = 25.62; Arm 3 – Auto-skip = 25.46; Arm 4 – Auto-play = 25.02). Thus, potentially harmful videos, and to a smaller extent neutral videos, were watched for longer in the intervention arms, compared to the control arm. (Note that the videos were trimmed to be engaging in the first 20-45 seconds of viewing, but the longest potentially harmful video was 55.27 seconds long).

Table 32. Mean view times of neutral videos and potentially harmful videos, by arm

Arm	Neutral videos mean total view time (in seconds)	Potentially harmful videos mean total view time (in seconds)
Arm 1 Control	24.33	23.72
Arm 2 Active choice	25.61	29.22
Arm 3 Auto-skip	25.46	29.53
Arm 4 Auto-play	25.02	28.63

7.7.6. Skips at the alert stage

Additional evidence that auto-skip alert messages increased the probability of participants skipping at the alert stage is shown in Figure 17. Specifically, Figure 17 shows the breakdown of when participants decided to skip potentially harmful videos across treatment arms. A higher proportion of participants in Arm 3 – Auto-skip (amongst those who eventually chose to skip potentially harmful videos) chose to skip at the alert stage (62%) rather than at the interface stage, relative to other those in the other intervention arms. Additionally, participants in Arm 4 – Auto-play were slightly less likely to skip potentially harmful content at the alert stage (24%) than those in the Arm 2 – Active choice (35%).

Figure 17. Proportion of skips that took place during the alert message stage relative to the interface stage, per intervention arm

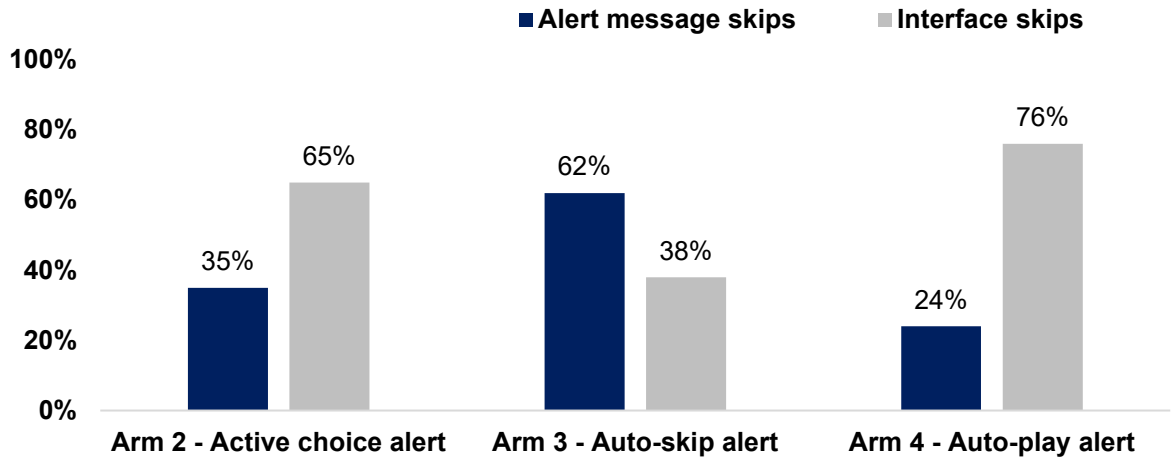
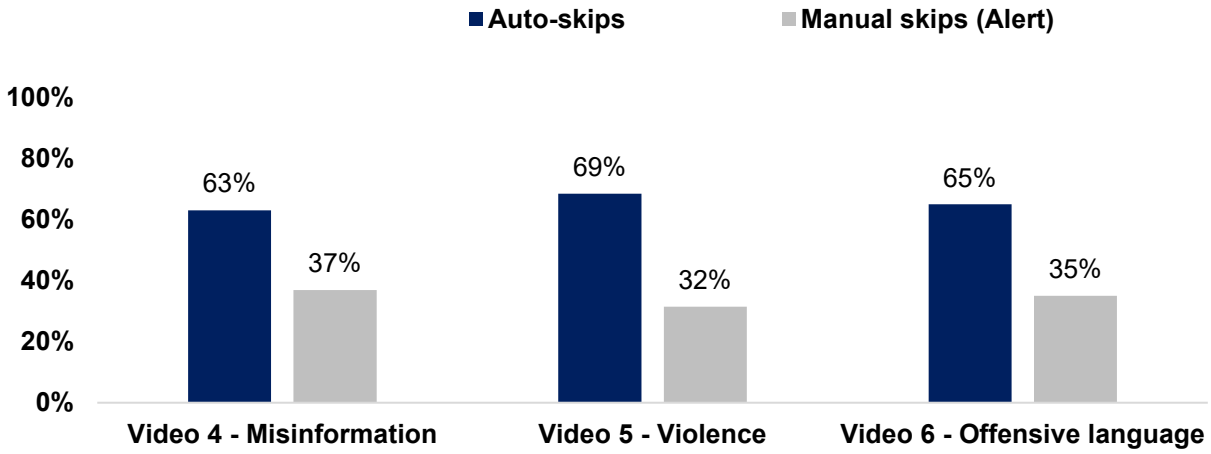


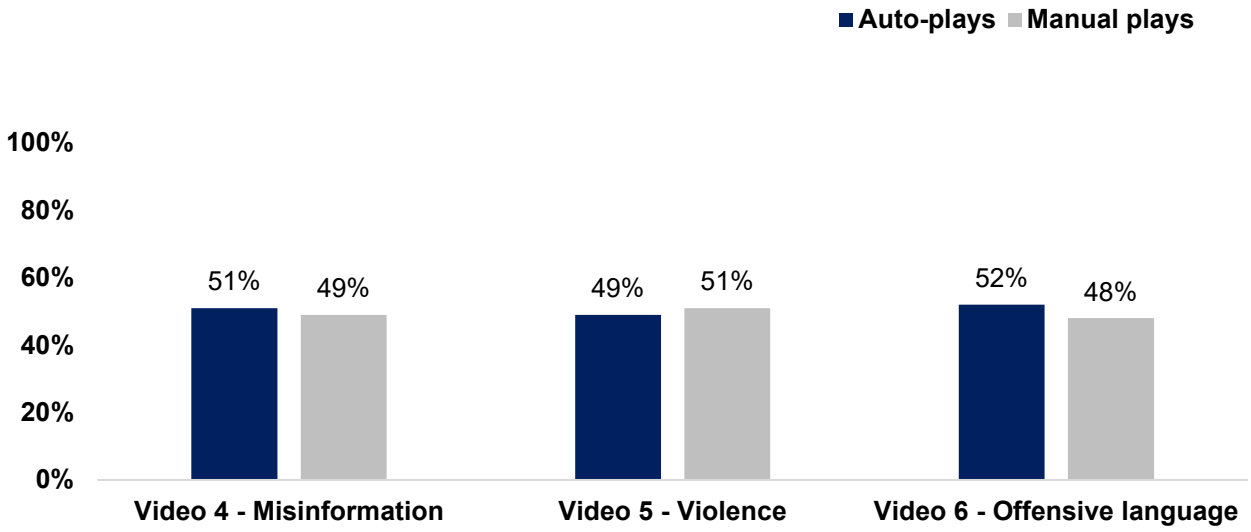
Figure 18 shows that amongst all skips made by participants exposed to auto-skip alerts at the alert stage, videos were most often skipped automatically via the countdown expiring (Video 4: 63%; Video 5: 69%; Video 6: 65%) rather than by manually skipping potentially harmful videos at the alert message. Thus, the reason for why the auto-skip intervention was successful may be that participants often passively waited for countdowns to finish and skip potentially harmful videos.

Figure 18. Proportion of skips that were manual or automatic at the auto-skip alert stage (Arm 3 – Auto-skip)



In contrast, Figure 19 shows that participants exposed to auto-play alerts were approximately equally likely to manually play or allow the auto-play to play potentially harmful videos at the alert stage. This suggests that participants exposed to auto-play alerts did not passively wait for potentially harmful videos to play to the same extent as participants exposed to auto-skip alerts passively waited for the potentially harmful videos to skip.

Figure 19: Proportion of plays that were manual or automatic at auto-play alert stage (Arm 4 Auto-play)



8. Comments

Adding alert messages with a countdown to auto-skip before each potentially harmful video significantly reduced the likelihood of users watching these videos, compared to any other arm (evidence for Hypotheses 1 and 3, section 4.1). This effect was robust to controlling for different types of devices that the experiment was completed on. Thus, we would expect it to generalise to the population of other panellists watching similar videos to the ones we had shown in the trial. Figure 17 in section 7.7.6 shows that participants exposed to auto-skip alerts were more likely to skip potentially harmful videos at the alert stage than those exposed to other alert messages. In addition, Figure 18 in section 7.7.6 shows that participants exposed to auto-skip alerts, were more likely to auto-skip than manually skip at the alert stage. Thus, the reason for why the auto-skip intervention was successful appears to be that participants were passively waiting for potentially harmful videos to be skipped. However, Figure 19 in section 7.7.6 shows that participants exposed to auto-play alerts were approximately equally likely to manually play or auto-play potentially harmful videos at the alert stage. If participants were passively waiting for a decision to be made for them at the alert stage, that would not explain why auto-play alerts were less successful at auto-playing potentially harmful videos than auto-skips were at auto-skipping potentially harmful videos. It may be that participants were more certain about wanting to play potentially harmful videos at the alert stage, than they were at wanting to skip them (given exposure to countdowns). Consequently, the countdown might have mattered less to participants in the auto-play arm. However, this is just speculation and more research would be needed to examine this potential causal pathway.

We found no robust differences in the probability of skipping potentially harmful videos between the other two alert messages used in this study and the control group without an alert message (no evidence to reject the null hypotheses of no difference). Critically, we found that none of the alert messages (which were shown before potentially harmful videos) had any effect on the probability of skipping of neutral videos. Consequently, the alert messages used in this study were not found to suppress engagement with neutral content (as measured using skips).

In addition to the findings concerning the probability of skipping potentially harmful videos, we found no differences in the probability of reporting potentially harmful videos between arms. However, we found significant differences in the amount of time spent watching potentially harmful videos. Specifically, participants who were exposed to alert messages, who chose to watch potentially harmful videos, watched them for longer than participants who were not exposed to alert messages.

Regarding responses to attitudinal questions, we found no significant differences in self-reported perceived usefulness, annoyingness, and distraction between participants exposed to different alert messages. Last, we also found no significant differences in self-reported levels of regret after choosing to watch potentially harmful videos, between participants exposed to different types of alert messages.

9. Appendix A: Post-trial questionnaire

Trial Arm (GROUP)	No. of questions	Format of questions
Control	2	Likert scale
Alert message only	11	Likert scale; multiple choice; Yes / No.
Alert message with Auto-play	12	Likert scale; multiple choice; Yes / No.
Alert message with Auto-skip	12	Likert scale; multiple choice; Yes / No.

COUNTDOWNCHECK

ASK IF GROUP = 3 OR 4

SINGLE CODE

In the warning messages you have just seen, was there a countdown to the video automatically:

1 Playing

2 Skipping

3 Not sure

At this point (before the next questions), for group 2, 3, 4

During the experiment, you were shown the following warning message before some of the videos:

SHOW STILL OF APPROPRIATE MESSAGE FOR GROUP 2, 3, 4

USEFUL

ASK IF GROUP = 2, 3, 4

SINGLE CODE

To what extent to you agree or disagree with the following statement:

'I found the warning messages useful when I was deciding whether to watch the videos.'

1 Disagree strongly

2 Disagree slightly

3 Neither agree nor disagree

4 Agree slightly

5 Agree strongly

6 Don't know

ANNOY

ASK IF GROUP = 2, 3, 4

SINGLE CODE

To what extent to you agree or disagree with the following statement:

'I found the warning messages annoying.'

1 Disagree strongly

2 Disagree slightly

3 Neither agree nor disagree

4 Agree slightly

5 Agree strongly

6 Don't know

ASK IF GROUP = 2, 3, 4

SINGLE CODE

To what extent do you agree or disagree with the following statement:

'I found the warning messages distracting.'

1 Disagree strongly

2 Disagree slightly

3 Neither agree nor disagree

4 Agree slightly

5 Agree strongly

6 Don't know

SCRIPTER NOTES: RANDOMISE THE ORDER OF USEFUL, ANNOY AND DISTRACT.

USEFULREASON2

ASK IF USEFUL = 4 "Agree slightly" OR 5 "Agree strongly" AND GROUP = 2

SINGLECODE

What did you find most useful about the warning messages you saw? Please select one.

1 The message made me stop and think about whether I really wanted to watch the video

2 The message forced me to make a choice to watch or to skip the video

3 The message covered the whole screen so I couldn't see the content before choosing

4 The message explained why I might find the content harmful

5 Other reason: [OPEN]

SCRIPTER NOTES: RANDOMISE OPTION ORDER FOR 1-4

USEFULREASON34

ASK IF USEFUL = 4 "Agree slightly" OR 5 "Agree strongly" AND GROUP = 3 OR 4

SINGLECODE What did you find most useful about the warning messages you saw? Please select one.

1 The message made me stop and think about whether I really wanted to watch it

2 The message forced me to make a choice to watch or to skip the video

3 The message covered the whole screen so I couldn't see the content before choosing

4 The message explained why I might find the content harmful

5 The message included a countdown

6 Other reason: [OPEN]

SCRIPTER NOTES: RANDOMISE OPTION ORDER FOR 1-5

NOUSEFULREASON

ASK IF USEFUL = 1 "Disagree strongly" OR 2 "Disagree slightly" OR 3 "Neither agree nor disagree"

SINGLECODE

Why didn't you find the warning messages useful or why were you unsure if they were useful? Please select one.

1 The warning message made me want to watch the video

2 I never read warning messages

3 I just wanted to get on and watch the content without being interrupted

4 I wouldn't find the content harmful

5 Other reason: [open]

SCRIPTER NOTES: RANDOMISE OPTION ORDER FOR 1-4

RUSH

ASK IF GROUP = 2, 3 OR 4

SINGLECODE

Did the warning messages make you feel rushed into making a choice about whether to watch or to skip the videos?

1 Yes

2 No

3 Not sure

SCRIPTER NOTES: RANDOMISE OPTION ORDER FOR 1-3

RECOMMENDATION

ASK IF GROUP = 2, 3 OR 4

SINGLECODE

Did you think that the warning messages were recommendations as to what you should do?

1 Yes

2 No

3 Not sure

SCRIPTER NOTES: RANDOMISE OPTION ORDER FOR 1-3

REGRET

ASK IF GROUP = 2, 3 OR 4 AND IF POTENTIALLY HARMFUL VIDEO VIEWED

FOR EACH POTENTIALLY HARMFUL VIDEO VIEWED

SINGLECODE

We noticed that you continued to watch the video after seeing a warning message. To what extent do you agree with the following statement?

'I regretted watching the video.'

1 Disagree strongly

2 Disagree slightly

3 Neither agree nor disagree

4 Agree slightly

5 Agree strongly

6 Don't know

SCRIPTER NOTES: DISPLAY SMALL SCREENSHOT FOR EACH POTENTIALLY HARMFUL VIDEO THAT PARTICIPANTS VIEWED ABOVE THE QUESTION

SCRIPTER NOTES: RANDOMISE THE ORDER OF RUSH, RECOMMENDATION AND REGRET.

ATTITUDE1

ASK ALL

SINGLECODE

You're now going to see some statements that other people have made about watching content on Video Sharing Platforms (VSPs). To what extent do you agree or disagree with the following statement:

'VSP users must be protected from watching potentially harmful content online.'

1 Disagree strongly

2 Disagree slightly

3 Neither agree nor disagree

4 Agree slightly

5 Agree strongly

6 Don’t know

ATTITUDE2

ASK ALL

SINGLECODE

To what extent do you agree or disagree with the following statement:

‘VSP users must be able choose what they watch online, even if it is potentially harmful.’

1 Disagree strongly

2 Disagree slightly

3 Neither agree nor disagree

4 Agree slightly

5 Agree strongly

6 Don’t know

SCRIPTER NOTES: RANDOMISE THE ORDER OF STATEMENTS IN ATTITUDE1 AND ATTITUDE2.

10. Appendix B: Demographic breakdown of participants by device type

Appendix B Item 1: Split of participants by device operating system and age group

Device operating system	18-24 (%)	25-39 (%)	40-54 (%)	55-69 (%)
Android	24.1%	41.0%	47.3%	33.4%
Ios	55.6%	30.3%	20.1%	11.7%
Windows	12.7%	22.7%	23.6%	39.0%
macOS	5.1%	3.8%	5.7%	7.5%
iPadOS	0.3%	0.2%	0.6%	3.4%
ChromeOS	1.3%	1.6%	2.0%	3.7%
Linux	0.0%	0.2%	0.6%	1.0%
Unknown	0.0%	0.2%	0.1%	0.3%

Appendix B Item 2: Split of participants by device operating system and gender

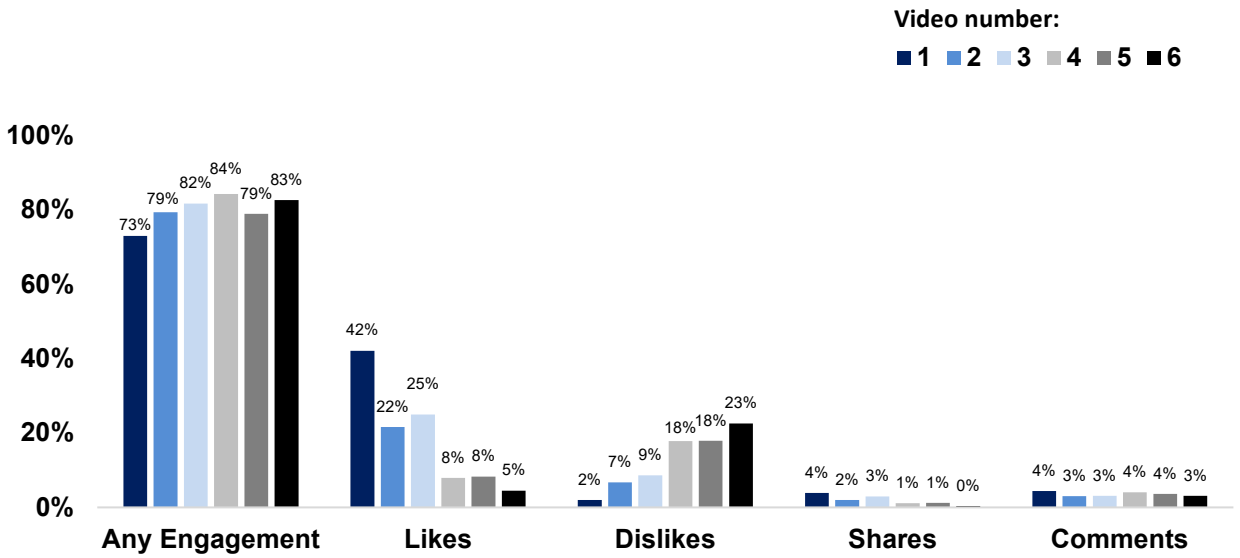
Device operating system	Male (%)	Female (%)	Other (%)	Prefer not to say (%)
Android	39.2%	38.6%	27.3%	33.3%
Ios	20.1%	33.2%	36.4%	33.3%
Windows	31.3%	19.2%	27.3%	33.3%
macOS	5.0%	5.7%	9.1%	0.0%
iPadOS	1.3%	0.8%	0.0%	0.0%
ChromeOS	2.0%	2.3%	0.0%	0.0%
Linux	0.8%	0.1%	0.0%	0.0%
Unknown	0.2%	0.2%	0.0%	0.0%

Appendix B Item 3: Split of participants by device operating system and SEG

Device operating system	ABC1 (%)	C2DE (%)
Android	34.2%	44.7%
Ios	25.7%	28.3%
Windows	30.4%	18.4%
macOS	5.9%	4.6%
iPadOS	1.1%	1.0%
ChromeOS	2.1%	2.2%
Linux	0.4%	0.6%
Unknown	0.2%	02%

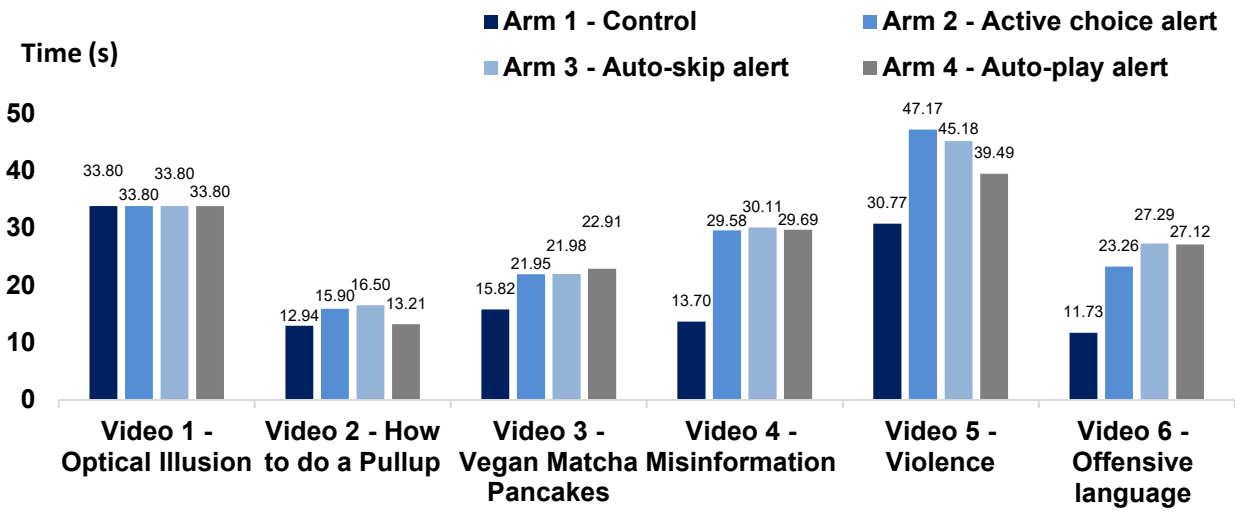
11. Appendix C: Additional descriptive statistics

Appendix C Item 1: Percentage of participants who engaged with each video

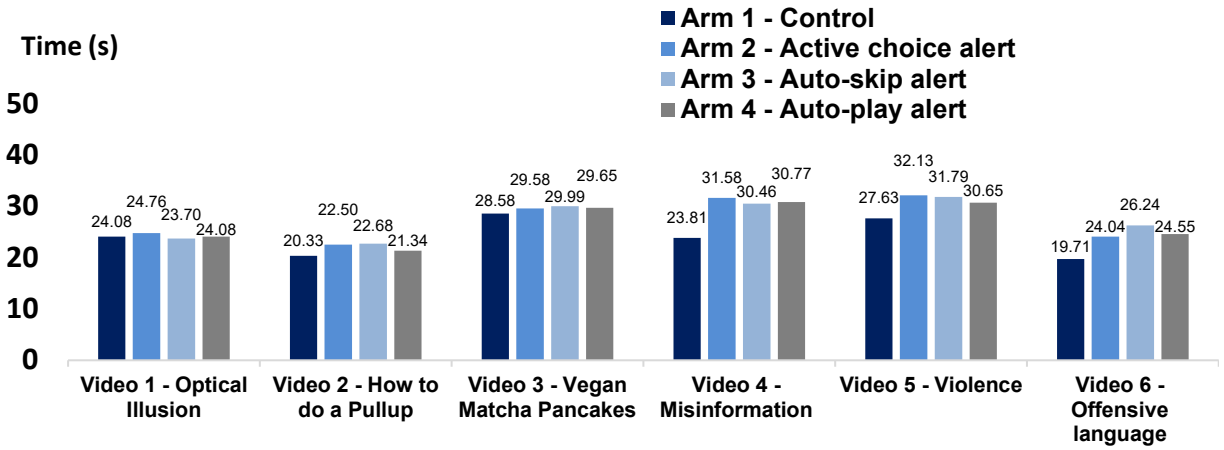


Note: Videos 1, 2, and 3 contained neutral content. neutral videos. Videos 4 (Misinformation), 5 (Violence) and 6 (offensive language) were categorised as potentially harmful.

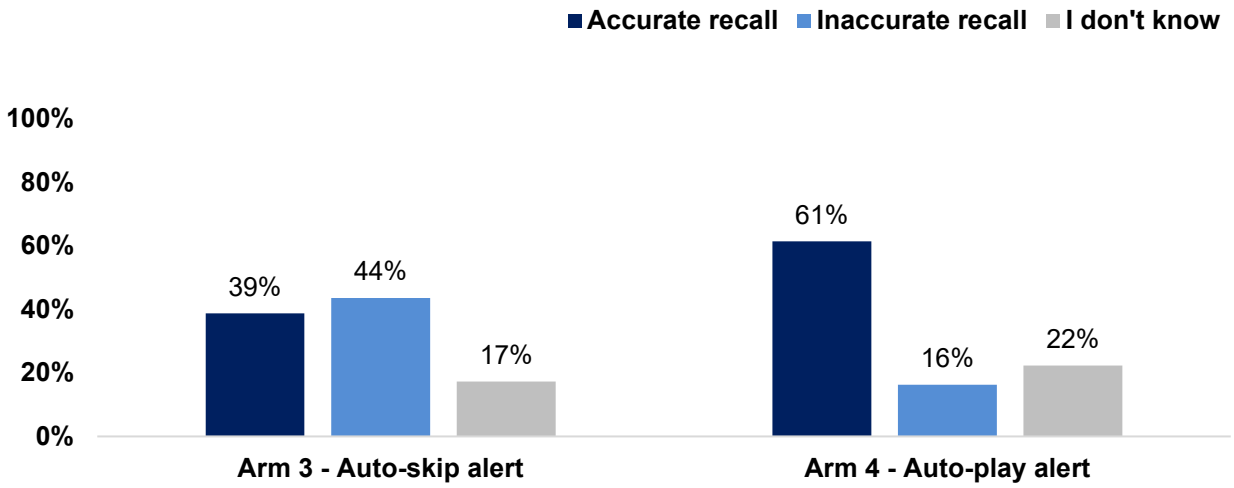
Appendix C Item 2: Average (median) viewing time of each video, by arm



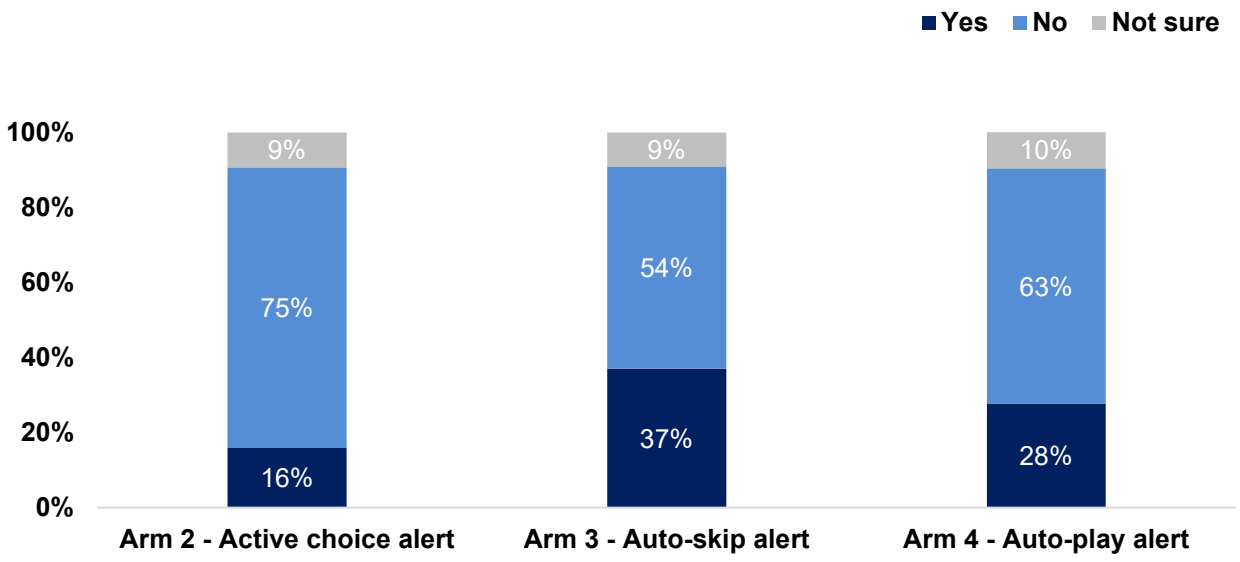
Appendix C Item 3: Average (mean) viewing time of each video split by arm



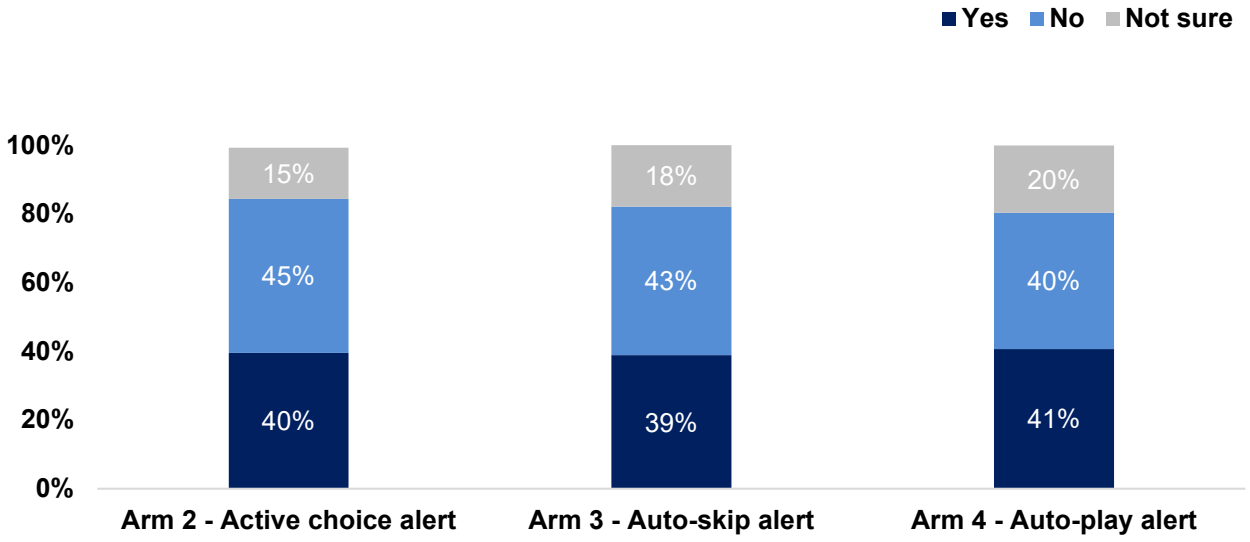
Appendix C Item 4: Percentage of participants accurately recalling the intervention they experienced



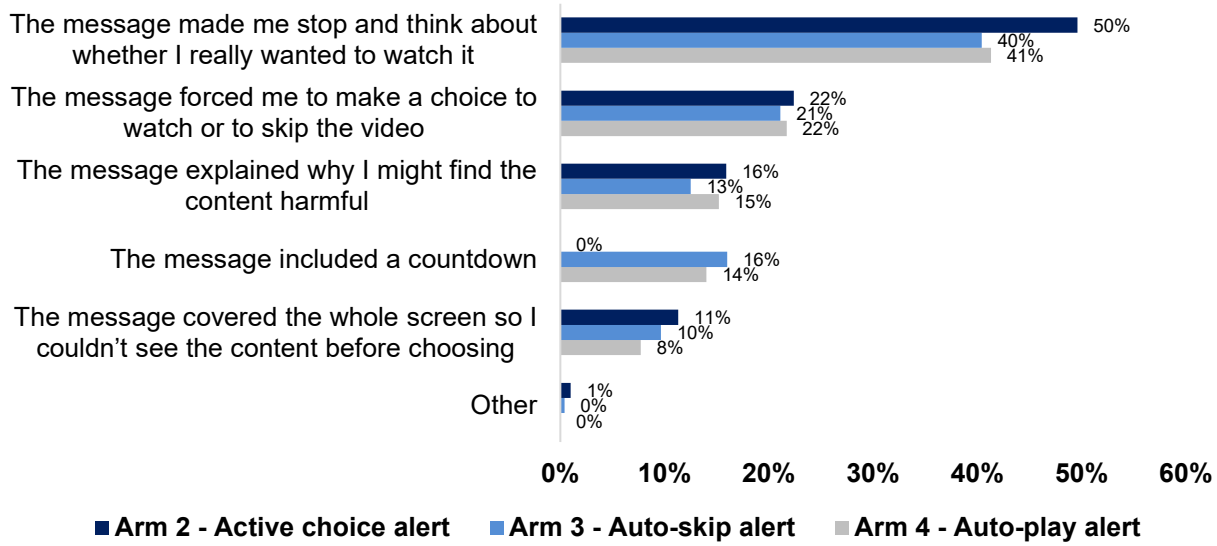
Appendix C Item 5: Percentage of participants who 'felt rushed into making a choice about whether to watch or skip videos', by arm



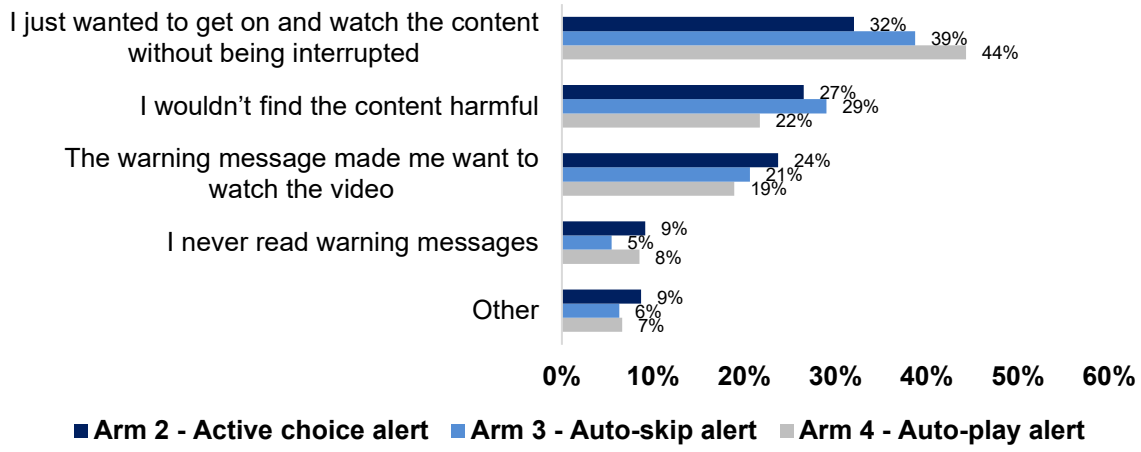
Appendix C Item 6: Percentage of participants who 'felt the warning messages were recommendations as to what [they] should do', by arm



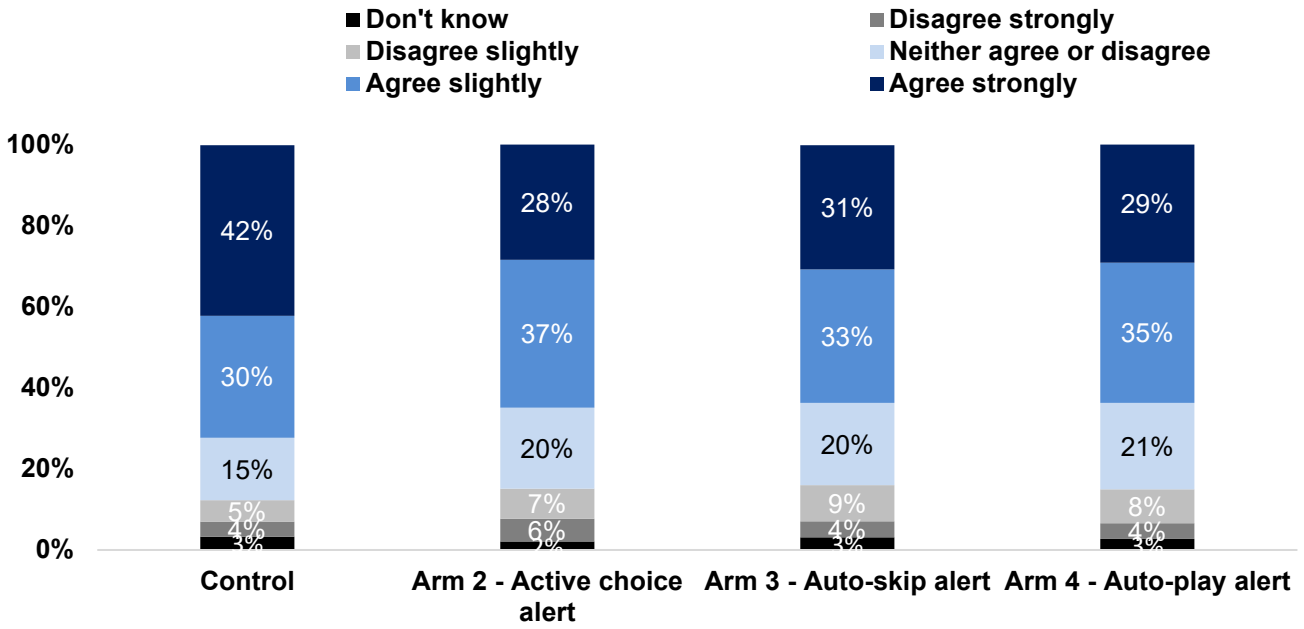
Appendix C Item 7: Reasons participants felt the alert message was useful, by arm



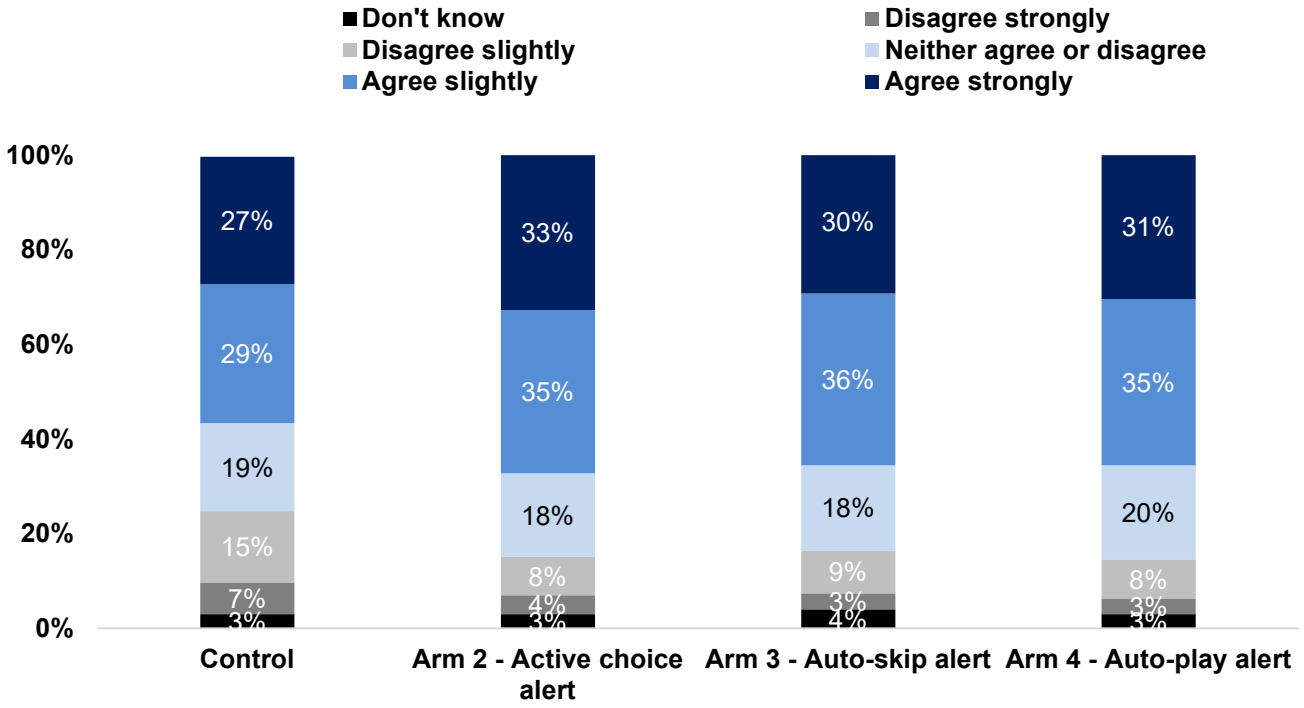
Appendix C Item 8: Reasons participants felt the alert message was NOT useful, by arm



Appendix C Item 9: Participant responses to the following prompt: 'VSP users must be protected from watching potentially harmful content online.', by arm



Appendix C Item 10: Participant responses to the following prompt: 'VSP users must be able choose what they watch online, even if it is potentially harmful.', by arm



KANTAR PUBLIC

Appendix C Item 11: Median response time to alert messages during the 1st, 2nd and 3rd encounter, by arm (excluding auto-skips and auto-plays)