Testing content controls to tackle online harms

Technical Report with methodology and results

Prepared by the Behavioural Insights Team and Ofcom

21 May 2024

THE
BEHAVIOURAL
INSIGHTS
TEAM

Contents

1. Background

1.1 Policy and regulatory context

Ofcom has a duty to promote and research media literacy and to carry out research into media literacy matters.[1] This includes user ability to control the content they see on their social media feed. Ofcom is also the regulator for video-sharing platforms (VSPs) and since November 2020, VSPs established in the UK must comply with rules around protecting users from harmful videos. Ofcom commissioned the Behavioural Insights Team (BIT) to run a randomised control trial (RCT) to test different interventions that empower users to align content with their preferences.

Additionally, this research will build evidence with respect to Ofcom's new duties under the Online Safety Act 2023 (OSA).[2] For example, this work is relevant to the user empowerment (UE) duties which require providers of category 1[3] services to offer users control features which reduce the likelihood of seeing certain types of content (described in section 16 of the OSA) at the earliest possible opportunity.

Finally, this work adds to the research Ofcom's Behavioural Insight Hub is carrying out to explore if and how platform design changes can be used to reduce online harms, such as the previous online trials on content reporting and alert messaging.[4]

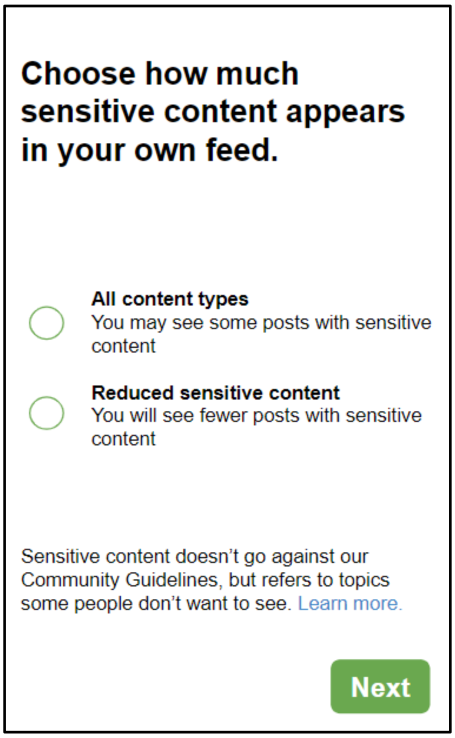
1.2 Research objectives

Understanding how online choice architecture can empower or disempower users to make decisions about the amount of sensitive content[5] on their feed is important for media literacy and online safety. The trial tested different ways of informing users during the sign-up for a social media platform of their options for controlling sensitive content (Figure 1). Such controls are referred to as content controls or content settings in this research. The interventions were designed to align users' initial choice of sensitive content controls with their preferences. To measure this alignment, we used the following approach. Participants made their initial choice at sign-up stage, saw the feed based on that choice and then were asked to confirm or change their choice ('Review' stage). Keeping the initial choice was interpreted as an indication that participants were able to make a well-informed initial choice

[1] UK Parliament, 2003. Communications Act 2003.
[2] UK Parliament, 2023. Online Safety Act 2023.
[3] Certain online services will be designated as a category 1, 2A or 2B services, depending on the number of users of the service, its functionalities, and any other relevant characteristics. The thresholds for each category will be set out in secondary legislation made by the Secretary of State.
[4] Ofcom, 2022. Behavioural insights for online safety: understanding the impact of video sharing platform (VSP) design on user behaviour.
[5] In this report, 'sensitive content' refers to content that is legal but that some users could find distressing or upsetting. For the full definition provided to research participants, see Figure 3.

that aligned with their preferences. We also asked participants about the reasons for keeping or changing their choice. See section 3.1 for further details about the trial design.

Figure 1. Example sensitive content settings page during a sign-up process.



Choose how much sensitive content appears in your own feed.

○ **All content types**
You may see some posts with sensitive content

○ **Reduced sensitive content**
You will see fewer posts with sensitive content

Sensitive content doesn't go against our Community Guidelines, but refers to topics some people don't want to see. Learn more.

Our interventions did not seek to steer users towards a particular choice (e.g. increase the number of adult users that choose “Reduced sensitive content”). Our primary outcome focused on the alignment between participants' choices and their preferences. As secondary outcomes, we also examined how the different interventions affected participants' comprehension of what type of content counts as sensitive and their overall sentiment towards content control settings (see section 3.5 for the full analytical framework). The trial aimed to answer the following main research questions:

RQ1: Does making information about the types of content categorised as sensitive more salient improve the alignment between participants' choices and their preferences?

RQ2: Does providing information about the types of content categorised as sensitive through an interactive microtutorial improve the alignment between participants' choices and their preferences?

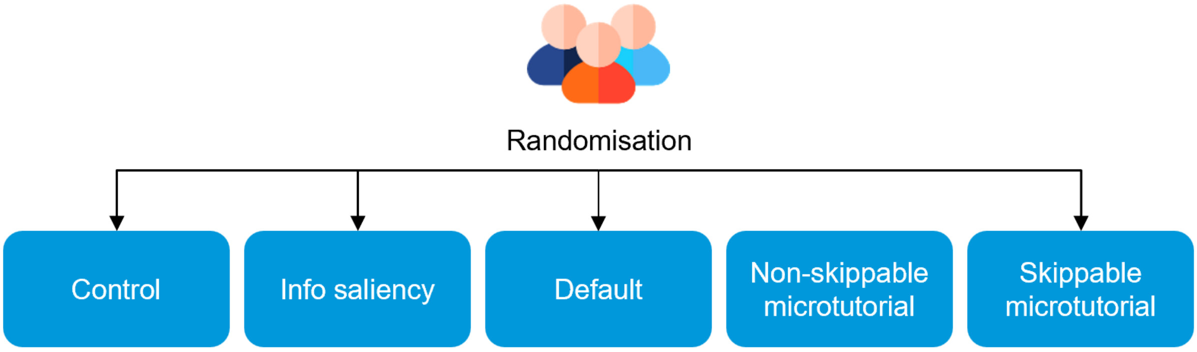
RQ3: How does defaulting participants into seeing “All content types” affect the alignment between participants' choices and their preferences?

2. Interventions and hypotheses

2.1 Trial arms overview

We conducted a five-arm RCT with one control and four treatment[6] conditions. Figure 2 gives an overview of the trial arms participants were randomised into.

Figure 2. Overview of trial arms.



2.2 Control arm

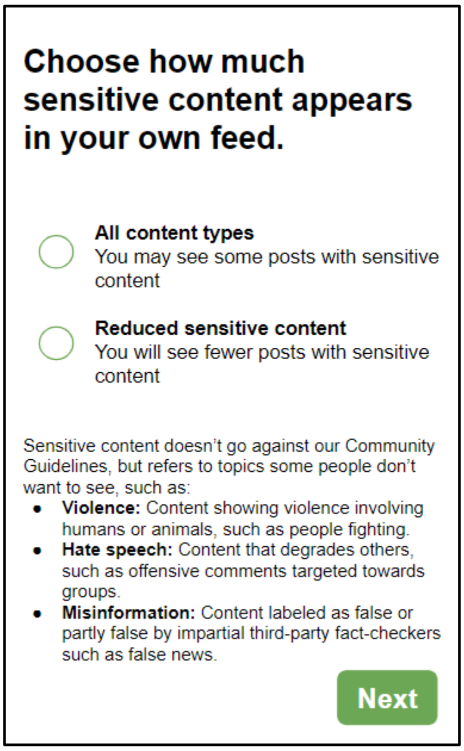
When designing the Control arm of the trial, we aimed to reflect the design popular platforms currently use for their sensitive content settings (see Figure 1). As is common practice on platforms, the settings page in the Control arm allows users to access a more elaborate definition of sensitive content by clicking ‘Learn more’.

2.2 Info saliency arm

The first intervention tested whether making the examples of sensitive content more salient and easier to access by having them on the settings page helps users make a more informed choice (see Figure 3). See section 3.4 for further information about how the content types were selected.

[6] Please note, we use the term ‘intervention’ and ‘treatment’ arms interchangeably in this report.

Figure 3. Info saliency intervention arm.[7]



Choose how much sensitive content appears in your own feed.

○ **All content types**
You may see some posts with sensitive content

○ **Reduced sensitive content**
You will see fewer posts with sensitive content

Sensitive content doesn't go against our Community Guidelines, but refers to topics some people don't want to see, such as:

- **Violence:** Content showing violence involving humans or animals, such as people fighting.
- **Hate speech:** Content that degrades others, such as offensive comments targeted towards groups.
- **Misinformation:** Content labeled as false or partly false by impartial third-party fact-checkers such as false news.

We expected that participants would be more likely to read these examples and understand the choice they were making. This, in turn, may affect the users' decisions and actions.

H1a: The probability of participants changing their content settings after having seen the feed will be significantly lower in the Info saliency arm compared to the Control.

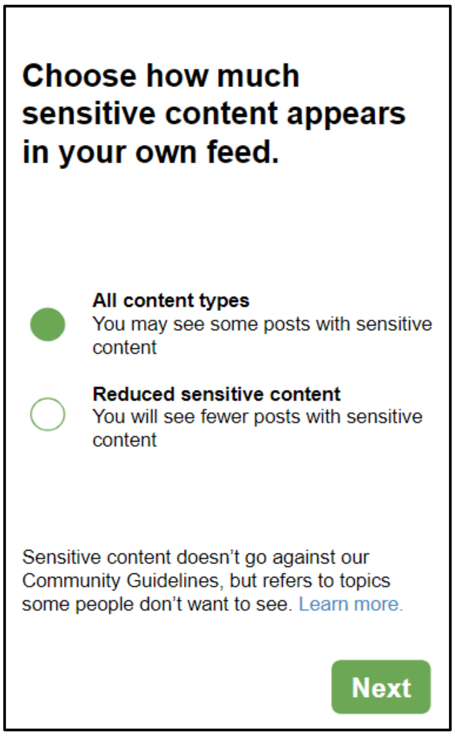
H1b: The probability of participants correctly identifying content as sensitive will be significantly higher in the Info saliency arm compared to the Control.

2.3 Default arm

The aim of this trial arm was to examine how pre-setting a default impacts the alignment of users' choices with their preferences compared to the Control arm without such pre-selection (see Figure 4).

[7] The descriptions of sensitive content included in the trial were for illustrative purposes only. They do not represent Ofcom's view on the description or definition of such categories of content for the purposes of the OSA.

Figure 4. Default setting intervention arm.



Choose how much sensitive content appears in your own feed.

● **All content types**
You may see some posts with sensitive content

○ **Reduced sensitive content**
You will see fewer posts with sensitive content

Sensitive content doesn't go against our Community Guidelines, but refers to topics some people don't want to see. Learn more.

We expected that pre-selecting the "All content types" option would increase the likelihood of participants initially choosing this option compared to the Control, but that after seeing the feed participants would be more likely to change their choice to "Reduced sensitive content".

H2a: The probability of participants changing their content settings after having seen the feed will be significantly higher in the Default arm compared to the Control.

H2b: The probability of participants correctly identifying content as sensitive will be significantly lower in the Default arm compared to the Control.[8]

H2c: The number of participants that initially choose to see "Reduced sensitive content" will be significantly lower in the Default arm compared to the Control.[9]

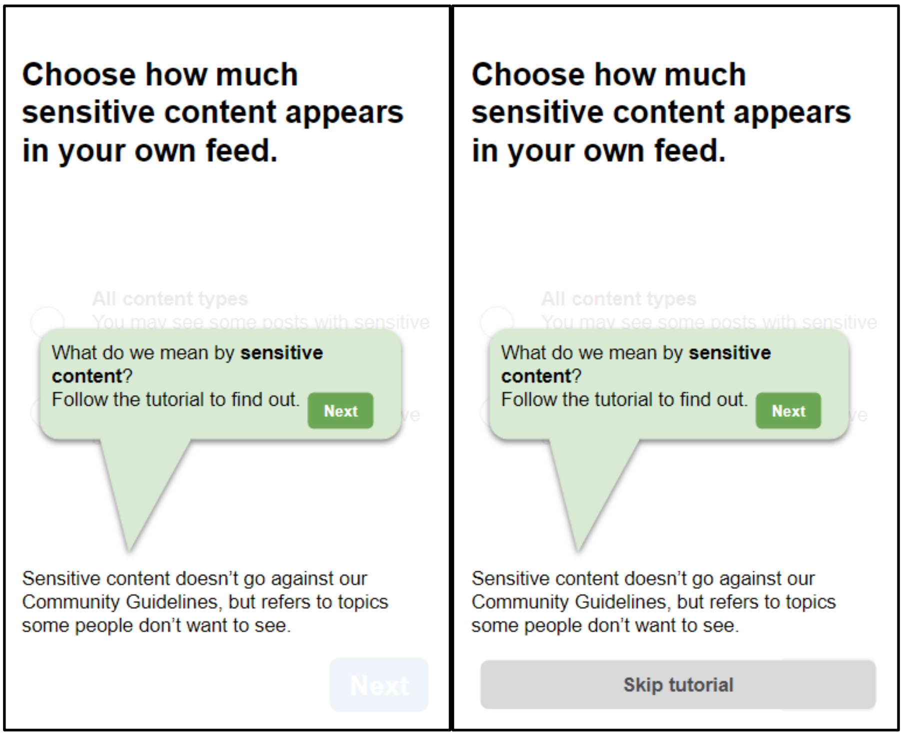
[8] Information provided to participants in Default and Control arms was the same. However, we expected that participants in the Default arm would be less motivated to engage with the information as one of the choices was pre-selected for them.

[9] This was part of exploratory analysis for which statistical testing did not involve adjustments for multiple comparisons. The results of the exploratory analysis should be treated with caution and do not allow us to formally conclude whether this hypothesis can be rejected or not.

2.4 Non-skippable and Skippable microtutorial arms

Microtutorials are short step-by-step guides designed to build capabilities in online behaviour. Unlike nudges that steer decisions[10], microtutorials aim to boost users' capabilities to make their own choices.[11] The trial tested whether walking participants through the sensitive content definitions as part of an interactive microtutorial helps align their setting choice with their content preferences. We included a non-skippable and skippable microtutorial arm as both types of design can be observed on actual online platforms (see Figure 5). All steps of the microtutorials are illustrated in Annex A.

Figure 5. First screen of the non-skippable and skippable interactive microtutorial.



Choose how much sensitive content appears in your own feed.

What do we mean by **sensitive content**? Follow the tutorial to find out.

Sensitive content doesn't go against our Community Guidelines, but refers to topics some people don't want to see.

Choose how much sensitive content appears in your own feed.

What do we mean by **sensitive content**? Follow the tutorial to find out.

Sensitive content doesn't go against our Community Guidelines, but refers to topics some people don't want to see.

All content types
You may see some posts with sensitive
Next
Next
All content types
You may see some posts with sensitive
Next
Skip tutorial

H3a: The probability of participants changing their content settings after having seen the feed will be significantly lower in the Non-skippable microtutorial arm compared to the Control.

H3b: The probability of participants correctly identifying content as sensitive will be significantly higher in the Non-skippable microtutorial arm compared to the Control.

H4a: The probability of participants changing their content settings after having seen the feed will be significantly lower in the Skippable microtutorial arm compared to the Control.

H4b: The probability of participants correctly identifying content as sensitive will be significantly higher in the Skippable microtutorial arm compared to the Control.

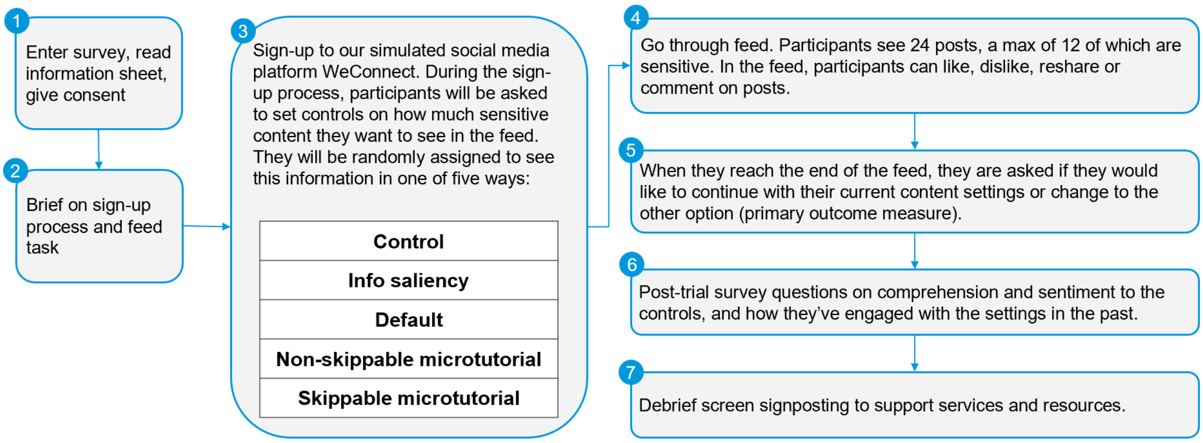
[10] Hertwig, R., & Grüne-Yanoff, T., 2017. Nudging and boosting: Steering or empowering good decisions, *Perspectives on Psychological Science*, 12(6), 973-986.
[11] Ofcom, 2023. Boosting users' safety online: Microtutorials.

3. Methodology

3.1 Trial design

To answer our research questions, we designed a simulated social media platform that mimicked real platforms. The simulated environment was embedded into an experimental survey with an RCT design. In an RCT, research participants are randomly divided into different groups and exposed to either an intervention arm or a control arm. Due to the random assignment into trial arms, intergroup differences in outcome measures can be causally attributed to the interventions participants were exposed to. Our trial design allowed us to measure the causal impact of the interventions in the sign-up process on participants' behaviours, decisions, and sentiment. Figure 6 illustrates the flow of the experiment.

Figure 6. Participant journey.



1 Enter survey, read information sheet, give consent

2 Brief on sign-up process and feed task

3 Sign-up to our simulated social media platform WeConnect. During the sign-up process, participants will be asked to set controls on how much sensitive content they want to see in the feed. They will be randomly assigned to see this information in one of five ways:

- **Control**
- **Info saliency**
- **Default**
- **Non-skippable microtutorial**
- **Skippable microtutorial**

4 Go through feed. Participants see 24 posts, a max of 12 of which are sensitive. In the feed, participants can like, dislike, reshare or comment on posts.

5 When they reach the end of the feed, they are asked if they would like to continue with their current content settings or change to the other option (primary outcome measure).

6 Post-trial survey questions on comprehension and sentiment to the controls, and how they've engaged with the settings in the past.

7 Debrief screen signposting to support services and resources.

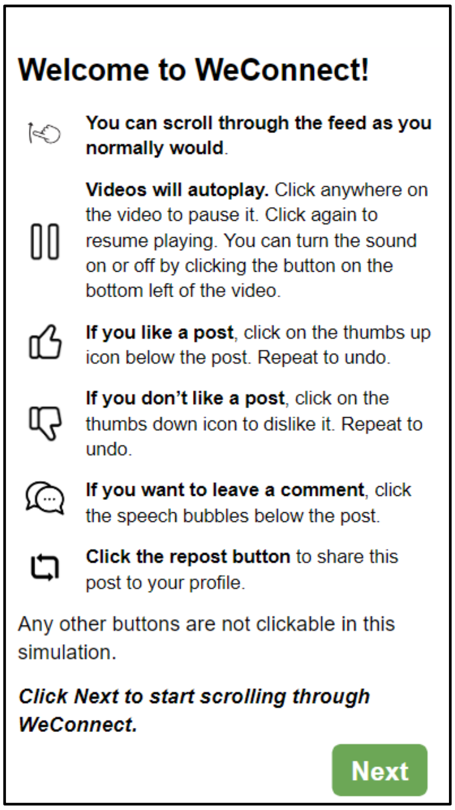
3.2 Simulated social media platform

3.2.1 Platform design and functionality

We designed our online platform, WeConnect, to create a trial environment that mimics real experiences on social media, increasing the external validity of our findings. External validity refers to the extent to which the findings of a study can be generalised to, and are representative of, real-world populations, settings, and conditions beyond the specific context of the research. While WeConnect is not based on any real-world platform, its design is inspired by popular platforms. By making participants' experiences on WeConnect as realistic as possible, we aimed to generate findings that indicate how our interventions would impact users' behaviours on real-world platforms.

The platform had two main components: 1) a sign-up process and 2) a content feed. During sign-up, participants went through a standard process where they were asked to allow push notifications, give their birthdate, decide on content settings and get introduced to the platform functionalities (see Figure 7).

Figure 7. Example screen from the sign-up process.



Welcome to WeConnect!

You can scroll through the feed as you normally would.

Videos will autoplay. Click anywhere on the video to pause it. Click again to resume playing. You can turn the sound on or off by clicking the button on the bottom left of the video.

If you like a post, click on the thumbs up icon below the post. Repeat to undo.

If you don't like a post, click on the thumbs down icon to dislike it. Repeat to undo.

If you want to leave a comment, click the speech bubbles below the post.

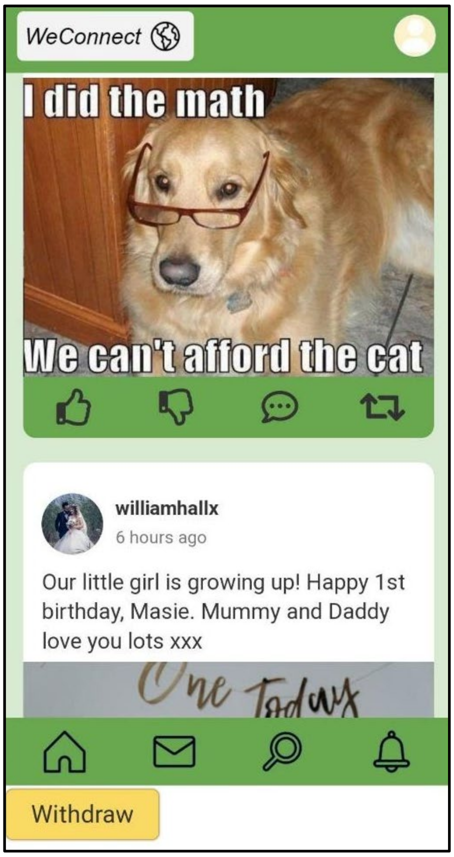
Click the repost button to share this post to your profile.

Any other buttons are not clickable in this simulation.

Click Next to start scrolling through WeConnect.

After the sign-up process, participants entered the content feed on WeConnect. Participants had to spend at least 60 seconds on the feed before they could progress to the next stage of the experiment. Participants could engage with the feed by liking, disliking, commenting and reposting posts. Figure 8 illustrates what the feed looked like. After participants scrolled through the feed and clicked 'Next' at the bottom, they progressed to the 'Review' stage, which asked whether they would like to change the content settings they chose when signing up to WeConnect (see section 3.5.1 for more detail).

Figure 8. WeConnect content feed.



williamhallx
6 hours ago

Our little girl is growing up! Happy 1st birthday, Masie. Mummy and Daddy love you lots xxx

3.2.2 Stimuli

Participants saw 24 pieces of content on their feed. The content consisted of 6 short videos, 6 long and 12 short text posts. Most of the text posts were accompanied by images related to the content of the post. The amount of content was informed by previous social media trials BIT ran and aimed to keep participants engaged in the feed for 5 minutes. Depending on the setting participants chose during the sign-up process, either 12 pieces of content (in the “All content types” setting) or 2 pieces of content (in the “Reduced sensitive content” setting) of the 24 pieces of content were sensitive. The sensitive content categories included in the trial are hate, violence, and misinformation (see section 3.4 for more details on content sourcing). The non-sensitive posts were made up of neutral content designed to resemble the type of content users encounter on real social media platforms. The content was presented on the feed in random order, apart from a few restrictions. To prevent an unrealistic scenario where participants would have to scroll through many potentially harmful items before encountering safe content again, we limited the exposure to sensitive content to no more than three pieces in a sequence. Furthermore, we chose to present participants with a non-sensitive post at the beginning and end of their feed. This approach aimed to balance the presentation of potentially distressing material with more neutral content.

3.2.3 Post-feed survey

After interacting with the main feed, participants completed a post-feed survey, which included questions on comprehension of what counted as sensitive content, their sentiment towards the content settings page, and their previous experience with content controls.

3.2.4 User testing

To ensure that our platform, the content, and the survey were understandable, easy to use and perceived as realistic, we conducted 7 user testing sessions with BIT employees not involved in the project. During these sessions, a BIT researcher worked closely with participants and had them think aloud (i.e. verbalise their thought processes) as they interacted with the experiment. Participants voiced their thoughts as they went through the platform and experiment, which gave us insight into their comprehension and areas of confusion. The researcher who led these sessions used a facilitation guide that included observation prompts on crucial aspects of the experimental design (e.g. does the user read the definition of sensitive content?).

BIT updated the design of the platform, interventions and survey questions based on user testing observations and feedback on the platform and content. Following the initial user testing, we updated the microtutorials to add more interactive features. This change was driven by observations that participants in the user tests were quickly clicking through the microtutorial content without engaging deeply with it. Further user testing with the revised interactive microtutorials was conducted to ensure these changes were perceived as intended. It indicated that these enhancements encouraged users to spend more time at each stage, which may have improved their understanding and retention of the material.

3.3 Sampling and data collection

3.3.1 Sample criteria

We recruited a nationally representative sample of adults from the UK. Participants were required to:

- be aged 18 years or older
- live in the UK

3.3.2 Power calculations

The sample size was based on power calculations for our primary outcome (whether the participant continued with their content settings choice). In the absence of published online experiments looking at comparable outcomes, we conducted calculations for baseline proportions ranging from 20%-50% (see Table 1), assuming 80% statistical power and a significance level of α = 1.25% (5% / 4; correcting for 4 comparisons in primary analyses).[12] A sample size of 3,500 participants (700 participants per arm) would allow us to detect a

[12] Note that for our analyses we use a Benjamini-Hochberg (BH) correction to adjust for multiple comparisons; however, it is not possible to apply this correction prior to data collection and so for power calculations we use a more conservative Bonferroni correction.

minimum detectable effect size of 8.88pp (percentage point difference) between a treatment arm and our Control arm where 50% of participants in the Control arm continued with their content settings choice. We deemed this sufficient for an online experiment and consistent with previous online experiments conducted by Ofcom.[13]

Table 1. Power calculations for a sample of 3,500 participants (700 per arm) assuming 80% statistical power and a significance level of α = 1.25%.

Outcome baseline	Minimum detectable effect size (percentage point difference)
20%	7.58pp
35%	8.71pp
50%	8.88pp

3.3.3 Data collection

All participants were recruited through the panel aggregator Lucid between 24 November and 14 December 2023. Participants spent an average of 7 minutes and 37 seconds completing the experiment. Each participant received financial compensation, with payments being administered by the panel providers they are registered with.[14]

To identify and mitigate any data protection risks, Ofcom and BIT conducted a data protection impact assessment of the research that was signed off by Ofcom. As part of the trial, no personal data was collected from the participants. BIT uses hashed IP addresses for online trials, meaning the data we collect is anonymised. Accordingly, it is impossible to identify which responses a particular participant provided. Participants were made aware of that through their panel providers before being redirected to our experiment.

To ensure there were no significant issues concerning data collection, we conducted a soft launch prior to the full launch of the trial. At this stage, the trial launched and recruited ~100 participants. Data collection was then paused while we conducted diagnostic checks to ensure data capture proceeded as planned and participants were not reporting any issues with the experiment. In the soft launch, we saw that the drop-out rate on the first page, before giving consent, and on the WeConnect page were high, so we updated the consent form to ask participants to click one tick box instead of two and specified in the trial that participants have to be in the feed for at least 60 seconds to continue to make this clearer for participants. There were no other data collection issues, so we proceeded to full launch. Soft launch data was used in the analysis. During data collection, we continued to monitor the

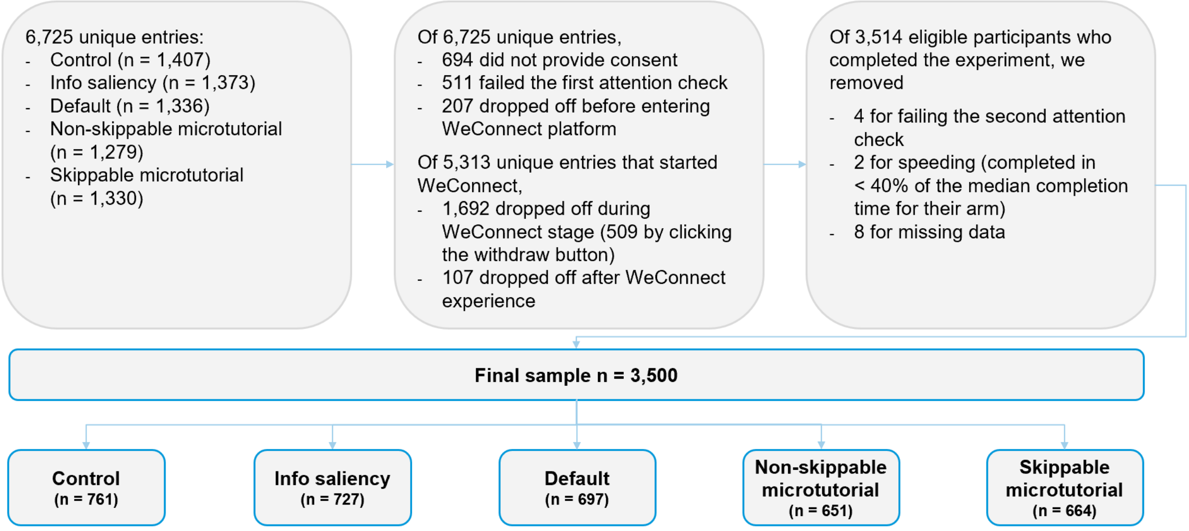
[13] Ofcom, 2023. Behavioural insights for online safety: understanding the impact of video sharing platform (VSP) design on user behaviour.

[14] The average compensation participants received was driven by the average time spent in the experiment and general market conditions. The average amount paid for participation in this trial reflected common payouts participants receive in similar online trials.

incoming sample against the quotas and flagged any criteria adjustments to the panel provider.

In the trial, we imposed additional pre-specified data quality measures in the form of attention and validation checks - only participants who passed these were retained for the analysis. The attention checks were brief questions near the beginning and the end of the trial, which asked participants to choose a particular response item to confirm they were paying attention. As a validation check, we looked at the time participants spent working through the trial and excluded those who were speeding through it (i.e. their survey completion time was less than 40% of the median completion time of that arm). Figure 9 shows the full participant flow with numbers on how many submissions were excluded at which part of the process.

Figure 9. Participant flow diagram



6,725 unique entries:
- Control (n = 1,407)
- Info saliency (n = 1,373)
- Default (n = 1,336)
- Non-skippable microtutorial (n = 1,279)
- Skippable microtutorial (n = 1,330)

Of 6,725 unique entries,
- 694 did not provide consent
- 511 failed the first attention check
- 207 dropped off before entering WeConnect platform

Of 5,313 unique entries that started WeConnect,
- 1,692 dropped off during WeConnect stage (509 by clicking the withdraw button)
- 107 dropped off after WeConnect experience

Of 3,514 eligible participants who completed the experiment, we removed
- 4 for failing the second attention check
- 2 for speeding (completed in < 40% of the median completion time for their arm)
- 8 for missing data

Final sample n = 3,500

Control (n = 761)

Info saliency (n = 727)

Default (n = 697)

Non-skippable microtutorial (n = 651)

Skippable microtutorial (n = 664)

3.4 Ethical considerations

The research went through BIT's and Ofcom's internal ethics review process and received full approval. The trial's main ethical and safeguarding concerns involved exposing participants, as well as BIT and Ofcom researchers, to sensitive content.

The UE duties apply to the types of content specified in section 16 of the OSA. Some of these content types (e.g. content related to suicide or self-harm) could not be included in the trial because of ethical considerations and the risk of causing serious harm to research participants. At the same time, it was necessary to expose participants to content that would go beyond neutral to generate evidence with high external validity.

Three categories of sensitive content were selected for the trial based on the following considerations:

- content being legal
- content types that could be considered potentially harmful but would not put participants at risk of serious harm

- content types used by Ofcom in previous research on VSPs

As a result, content types displaying hate, violence and misinformation were included in the trial. These do not directly correspond to the types of content specified in section 16 of the OSA but we considered they represent a broad range of sensitive content.

All text and imagery shown to participants in the trial were sourced from publicly available and freely reusable content (uploaded under a Creative Commons License) on platforms like YouTube and Unsplash. The age classification of all sensitive content was 18+, according to the BBFC content guidelines.[15]

The following risk mitigation and safeguarding measures were implemented to ensure the research did not cause harm to participants and researchers.

1. All content shown to participants in the trial has been reviewed and approved by BIT's ethics reviewer.

2. Participants could only access the trial if they agreed to consent forms provided to them beforehand. The consent forms detailed the research purpose and themes of the sensitive content. They outlined the potential risks involved in participating in the trial, so that participants, particularly those with specific vulnerabilities that might be triggered by the content included, could make an informed choice as to whether to participate. The consent form also made clear to participants that they could leave the survey at any moment without giving a reason.

3. The simulated platform included a visible 'Withdraw' button in the interface that made it easy to leave the trial immediately. Leaving the trial through this emergency button did not impact participants' eligibility for compensation.

4. Regardless of whether the participants decided to complete the study, a debriefing screen was provided with telephone numbers and links signposting to immediate support resources such as the Mind Infoline or the Samaritans hotline.

5. BIT staff voluntarily joined the research after a risk briefing and were allowed to withdraw at any point without penalties. If team members became distressed, they were allowed to switch to lower-risk roles.

6. Mental health support from BIT was available to the researchers, including Mental Health First Aiders and an Employee Assistance Programme. Ofcom equally implemented internal safeguards to protect staff exposed to sensitive content as part of this research.

7. When sensitive content was shared with Ofcom (e.g. for test-link preview), sensitive content warnings were used to alert staff involved in the trial to potential risks.

[15] BBFC. (n.d.). BBFC: View what's right for you [accessed 27 February 2024].

8. Ofcom equally implemented internal safeguards to protect staff exposed to sensitive content as part of this research.

3.5 Analytical framework

3.5.1 Data checks

First, we checked for differential attrition on a data set of participants who consented, passed the attention check and who made it to or past the WeConnect platform (*n* = 5,313; see Figure 9) using a linear regression with the last page of the experiment they completed as the outcome variable and the treatment arm as the predictor variable.

We then checked that our final sample (*n* = 3,500) was balanced in terms of demographics (age, gender, ethnicity, annual household income (pre-tax), education, urbanicity, employment, region) across treatment arms using chi-squared tests for categorical variables and analysis of variance for continuous variables.

3.5.2 Analytical strategy

We followed a pre-specified analysis framework which involved allocating our variables to primary, secondary, and exploratory outcomes based on an agreed upon hierarchy.

For outcomes with binary data (primary outcome and exploratory analyses 1-6) we conducted logit regressions and for outcomes with count data (secondary outcomes 1-2) we conducted Poisson regressions. For all models, our predictor variable was the treatment variable with the Control arm as the baseline, and we included age, gender, income, education, ethnicity, and platform use as covariates. We used a significance level of 5% throughout, correcting for 4 comparisons across the primary outcome and 8 comparisons across secondary outcomes using the Benjamini-Hochberg adjustment (no adjustments were made for exploratory analyses). It is BIT’s standard practice to use the Benjamini-Hochberg adjustment and correct for primary and secondary outcomes separately; however, this is different from the pre-specified analysis plan, where we said we would correct for 12 comparisons across primary and secondary comparisons using the Bonferroni adjustment. As the results of the primary and secondary analyses are null, using a less conservative correction did not make a difference to our analysis.

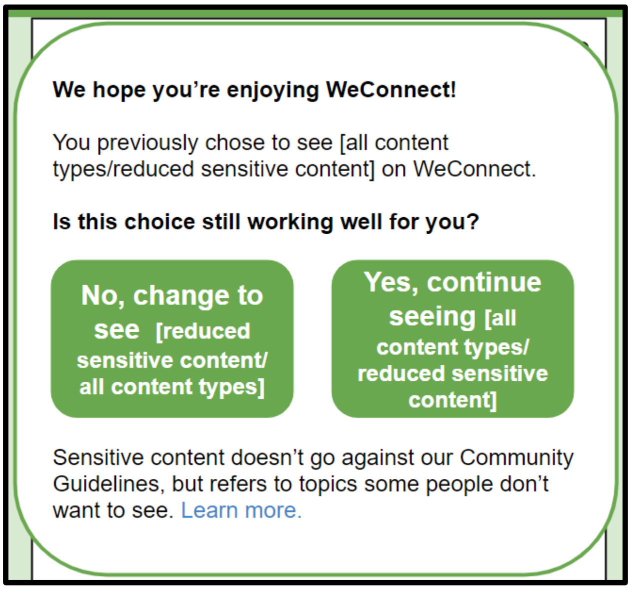
For the primary analysis, we checked the goodness of fit of our model using an ROC curve (receiver operating characteristic curve) plot and the AUC (area under the curve). The ROC curve represents the trade-off between the model's sensitivity (or true positive rate) and 1-specificity (false positive rate) at various threshold settings. A ROC curve that falls close to the top left corner of the plot (and therefore with a high AUC) indicates the model has high predictive performance. This can be compared against a random classifier model with predictive performance at chance level (and therefore a .5 AUC), where the true positive rate and false positive rate are the same. The ROC curve and AUC statistic can be used to evaluate the ability of a model to predict binary outcomes.

For the secondary analysis on overall sentiment score, we checked the data for any excess number of zeros or over-dispersed values. For exploratory outcomes 3-6, we conducted a sensitivity check and conducted ordinal regressions with the outcomes coded an ordinal rather than binary ("Not at all" coded as 0, "A little" coded as 1, "Moderately" coded as 2, and "Very much" coded as 3). We also checked for proportional odds of the ordinal regressions through a Brant test. Exploratory analyses are not corrected for multiple comparisons and results should not be taken as hypothesis confirming.

3.5.3 Primary analysis

After interacting with the feed, participants were asked whether their pre-feed content settings choice (to see all content types vs. reduced sensitive content) was working well for them (see Figure 10). In the pre-specified analysis plan, the primary outcome was formulated as whether participants chose to change their content controls after seeing the feed (change coded as 1, continue with initial choice coded as 0). However, on reflection, we decided that reporting the proportion of participants who chose to continue instead would be more consistent with our reporting and potential recommendations. Thus, we chose to revise our approach and code 1 as the choice to continue and code 0 as the choice to change the initial content settings. As this was just reversing the coding, this does not affect the analysis. In this report, we present the results as the proportion of people who chose to continue with their initial choice as the interventions (except for the Default arm) aimed to increase this proportion.

Figure 10. Primary outcome



We hope you're enjoying WeConnect!

You previously chose to see [all content types/reduced sensitive content] on WeConnect.

Is this choice still working well for you?

No, change to see [reduced sensitive content/ all content types]
Yes, continue seeing [all content types/ reduced sensitive content]

Sensitive content doesn't go against our Community Guidelines, but refers to topics some people don't want to see. Learn more.

3.5.4 Secondary analyses

Secondary outcome 1: After interacting with the feed, participants were asked a comprehension question which involved categorising 8 descriptions of posts (e.g. "a video of teenagers fighting on a playground") as either sensitive content or not sensitive content.

Secondary outcome 1 measured the number of content types that participants correctly categorised. Scores ranged from 0 to 8, with higher scores representing more correct categorisations.

Secondary outcome 2: After interacting with the feed, participants were asked about their sentiment towards the content settings page. They were asked whether they thought the content settings page was easy to understand, made them feel in control of the content they saw, was presented in a fair way, and whether they trusted that the choices were presented with their best interests in mind. For the four sentiment questions, participants could answer “Not at all” (coded as 0), “A little” (coded as 1), “Moderately” (coded as 2), and “Very much” (coded as 3). Responses to the four questions were summed, so overall sentiment scores could range from 0 to 12, with high scores reflecting more positive sentiments to the content settings page.

3.5.5 Exploratory analyses

Exploratory outcome 1: In the sign-up page, all participants were given the option of choosing to see either reduced sensitive content (coded as 1) or all content types (coded as 0).

Exploratory outcome 2: After interacting with the feed, participants were asked whether they would expect to see more sensitive content by selecting “All content types” vs. “Reduced sensitive content”, including an option to select “I don’t know”. We coded correct responses as 1 (“All content types”) and any other answers as 0.

Exploratory outcomes 3-6: To further explore secondary outcome 2, we analysed the post-feed sentiment questions separately. We looked at what the participants thought about the content settings in terms of: ease of understanding; feeling of control; presented in a fair way; and trust they were presented with their best interests in mind. For all outcomes: “Not at all” and “A little” coded as 0; “Moderately” and “Very much” coded as 1.

4. Results

4.1 Sample characteristics

We did not find evidence for an overall effect of differential attrition (adjusted $R^2 = 0.0006$, $F(4, 5308) = 1.86$, $p = .115$); however, there was some evidence indicating that participants in the Skippable microtutorial arm were more likely to drop off the experiment compared to the Control arm ($p = .038$). The demographics for our final sample ($n = 3{,}500$) are reported in Table 2. The sample was balanced across treatment arms for all variables (all $p > .05$), except for education, (X^2 (4) = 18.50, $p < .001$; % who are at least degree level educated: Control = 70%, Info saliency = 61%, Default = 63%, Non-skippable microtutorial = 69%, Skippable microtutorial = 68%). Despite this, by including education as covariate in all statistical models as planned, the effect of this imbalance is minimal. Given the sample was generally balanced across treatment arms, we continued with our pre-specified analysis plan.

Table 2. Sample demographics for final sample (n = 3,500)

	Category	% of the sample
Age	18-24	12%
	25-54	60%
	55 and over	29%
Gender	Male	47%
	Female	52%
	Other (e.g. non binary)	1%
Ethnicity	White	85%
	Asian	6%
	Black	6%
	Mixed or other	3%
Annual pre-tax income	£40,000 or over	47%
	Less than £40,000	53%
Education	Degree	31%
	No degree	66%
	Prefer not to say	3%
Urbanicity	Urban	29%
	Suburban	49%
	Rural	22%
Employed	Employed	68%
	Unemployed	3%
	Inactive	29%
Location	London	12%
	Midlands	16%
	North	26%
	South & East	32%
	Wales, Scotland & Northern Ireland	14%
Urbanicity	Urban	18%
	Suburban	50%
	Rural	31%
Social grade	High	34%
	Medium	57%

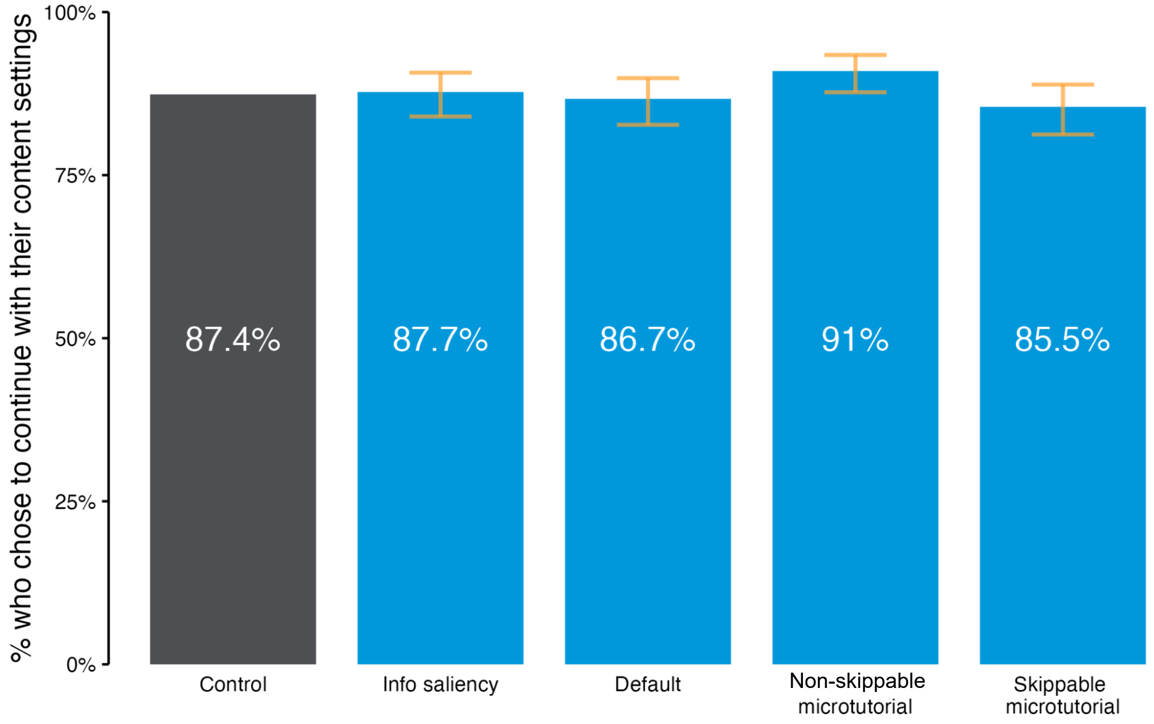
	Low	8%
	Don’t know	< 1%
Use social media	Yes	91%
	No	9%

Note. Some variables do not sum to 100% due to rounding.

4.2 Primary analysis: Whether participants continue with their initial choice

Overall, 88% chose to continue with the choice they made at sign-up, after browsing through WeConnect. After correcting for multiple comparisons, we found no significant differences when comparing the four intervention arms against the Control arm (p > .05). The results of the primary analysis are presented in Figure 11. The ROC curve for the primary model is presented in Figure 12. The diagnostic performance of the primary model was poor but better than chance (AUC = 0.60, bootstrapped 95% confidence intervals 0.58-0.63). Therefore, the primary analysis model was not good at predicting whether people continued with their content settings choice. However, the purpose of the model was to determine the causal effect of the treatments on the primary outcome, rather than classification, and so our main interpretation of the primary analysis holds — people in the treatment arms were not more likely to continue with their content settings choice than the Control arm.

Figure 11. The results of the primary analysis, comparing the percentage of participants who chose to continue with their content settings in the Control arm to each intervention arm.



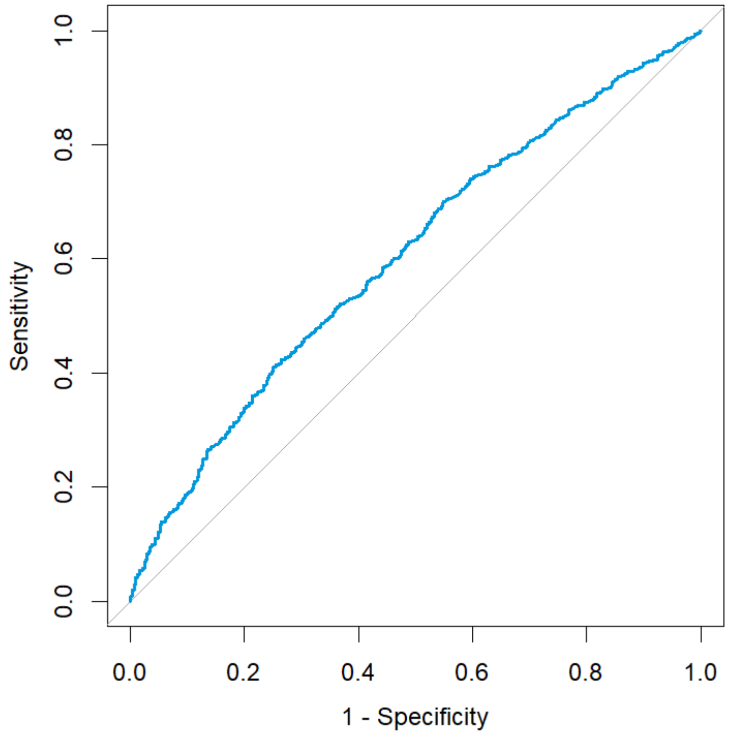
** p < .01, * p < .05, + p < .1

Significance is corrected for multiple comparisons. Confidence intervals (95%) are not corrected for multiple comparisons.

Regression controls for age, gender, income, ethnicity and platform use.

Treatment bars show the mean of the control + the coefficient of the respective treatment variable.

Figure 12. The ROC (receiver operating characteristic) curve for the primary analysis model



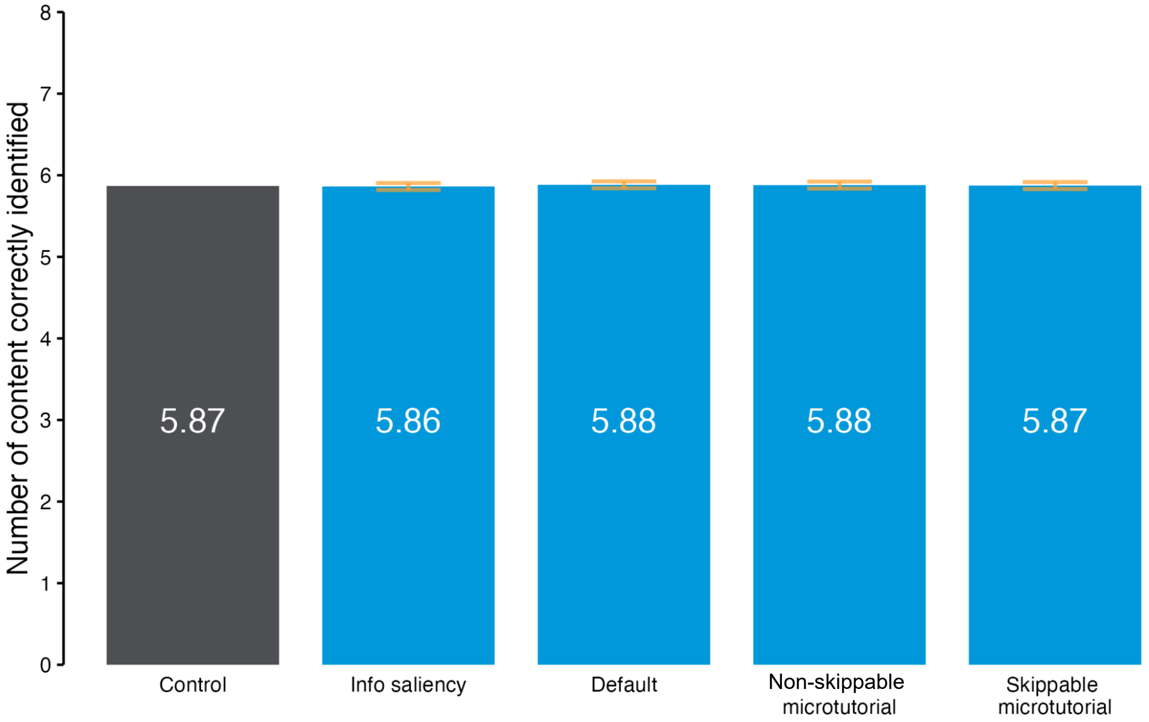
Note. The ROC curve (in blue) represents the trade-off between the model's sensitivity (or true positive rate) and 1-specificity (false positive rate) at various threshold settings. A ROC curve that falls close to the top left corner of the plot (and therefore with a high AUC) indicates the model has high predictive performance. This can be compared against a random classifier model (in grey) with predictive performance at chance level (and therefore a .5 AUC), where the true positive rate and false positive rate are the same.

4.3 Secondary analyses

4.3.1 Comprehension

Out of the 8 content descriptions participants were asked to identify as sensitive or not sensitive, participants correctly categorised on average 5.87 pieces of content. This was not significantly different between the Control and any of the treatment arms, *p* > .05. Results are shown in Figure 13.

Figure 13. The results of secondary analysis 1, comparing the content participants correctly identified as sensitive or not sensitive in the Control arm to each intervention arm.



** p < .01, * p < .05, + p < .1

Significance is corrected for multiple comparisons. Confidence intervals (95%) are not corrected for multiple comparisons.

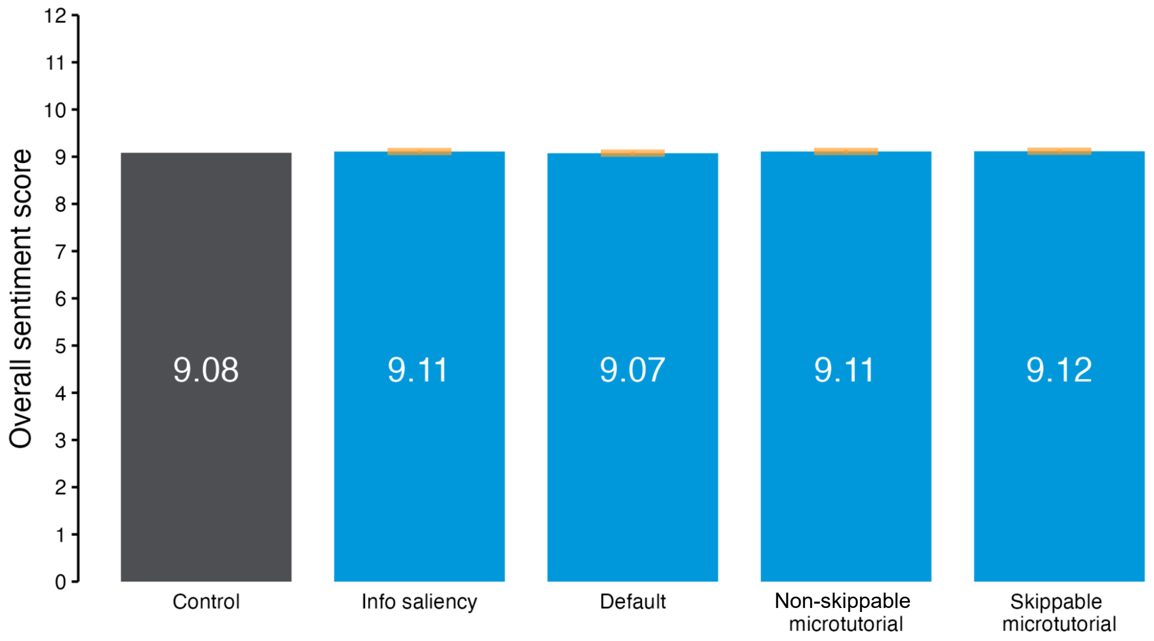
Regression controls for age, gender, income, ethnicity and platform use.

Treatment bars show the mean of the control + the coefficient of the respective treatment variable.

4.3.2 Sentiment

Before conducting secondary analysis 2, we checked the responses for an excess of zeros in the overall sentiment score. Only 7 participants had an overall sentiment score of 0. The values were only marginally under dispersed (X^2 (3484) = 2489.07, dispersion = 0.71), and so a normal Poisson regression was used as planned. The overall sentiment score was 9.10 out of a maximum of 12. This was not significantly different between the Control and any of the treatment arms, *p* > .05. Results are shown in Figure 14.

Figure 14. The results of secondary analysis 2, comparing the average overall sentiment score in the Control arm to each intervention arm.



** p < .01, * p < .05, + p < .1

Significance is corrected for multiple comparisons. Confidence intervals (95%) are not corrected for multiple comparisons.

Regression controls for age, gender, income, ethnicity and platform use.

Treatment bars show the mean of the control + the coefficient of the respective treatment variable.

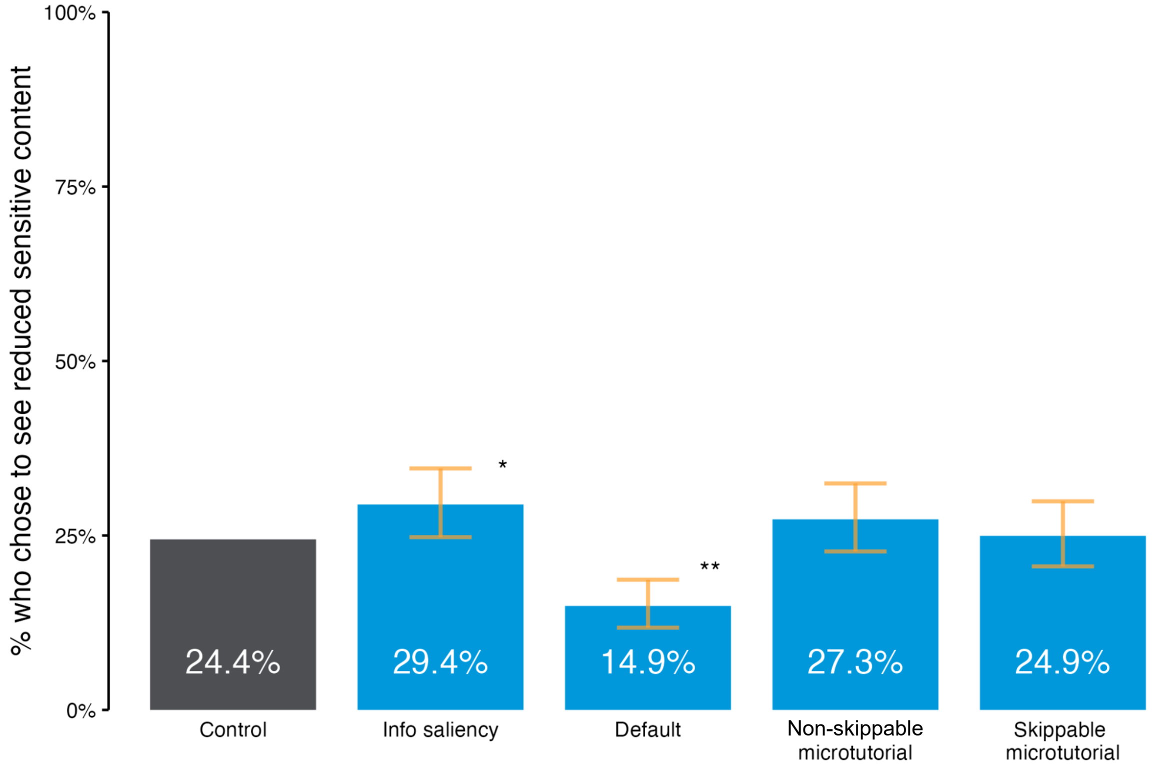
4.4 Exploratory analyses

Note that exploratory analyses have not been corrected for multiple comparisons. The approach to not do multiple comparison corrections for exploratory comparisons is driven by interpretation considerations. For exploratory comparisons we focus more on the direction and magnitude of effects, rather than significance and power. A significant result for an exploratory comparison is generally reported as an opportunity for further research. Exploratory comparisons help us to explain the results arising from our primary and secondary analyses, but they are not the focus of the interventions. This approach allows us to probe our primary and secondary results with exploratory analyses without attributing too much weight to false positive findings that can arise from a high number of comparisons. Correcting for multiple comparisons is a statistical adjustment made when analysing data that helps to reduce the probability of incorrectly rejecting a true null hypothesis (a ‘false positive’). Therefore, the findings in this section should be taken as exploratory rather than hypothesis confirming.

4.4.1 Choice

Overall, 24% of participants chose to see reduced sensitive content at sign-up.[16] Compared to the Control arm where 24.4% of participants chose to see reduced sensitive content, the info saliency led to a significant increase in participants choosing to see reduced sensitive content (29.4%, *p* < .05) and the default led to a significant decrease in participants choosing to see reduced sensitive content (14.9%, *p* < .01). There were no significant differences between the Control arm and the Non-skippable microtutorial arm or the Skippable microtutorial arm, *p* < .05. Results are shown in Figure 15.

Figure 15. The results of exploratory analysis 1, comparing the percentage of participants who chose to see reduced sensitive content in the Control arm to each intervention arm.



** p < .01, * p < 0.5, + p < .1

Significance and confidence intervals (95%) are not corrected for multiple comparisons.

Treatment bars show the mean of the control + the coefficient of the respective treatment variable.

Regression controls for age, gender, income, ethnicity and platform use.

4.4.2 Comprehension

Overall, 75% of participants correctly understood that the “All content types” option shows the most sensitive content on their feed. This was not significantly different between the

[16] At the end of the trial, 30.4% of participants had the “Reduced sensitive content” setting (30.5% in Control, 35.1% in Info saliency, 24.7% in Default, 31.0% in Non-skippable microtutorial and 30.6% in Skippable microtutorial)

Control and any of the treatment arms, *p* < .05. 5% said they do not know which option would show them the most sensitive content.

4.4.3 Sentiment

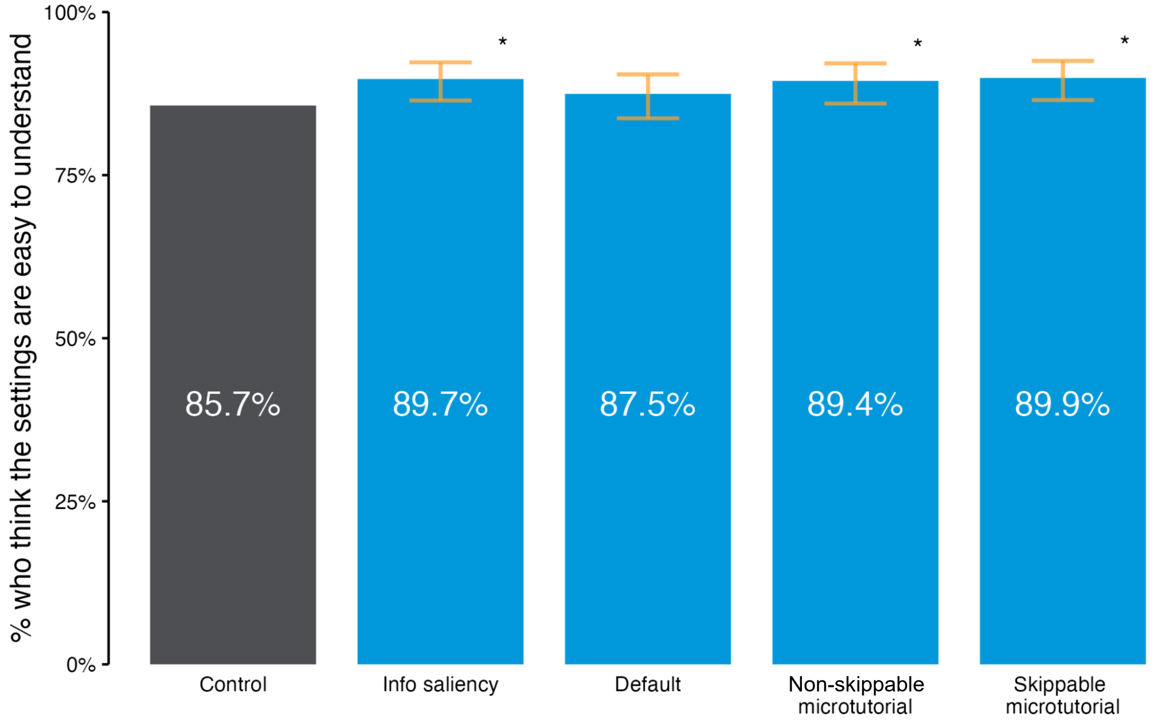
Results from ordinal logistic regressions and logistic regressions on binarised outcomes match on all exploratory sentiment outcomes besides ‘ease of understanding’. In this section, we report the results of the logistic regressions as well as the results of the ordinal logistic regressions for the ‘ease of understanding’ outcome. Results of all other ordinal logistic regressions are in the appendix.

Ease of understanding

85.7% of participants said they thought the content settings were moderately or very easy to understand. This was significantly higher for those in the Info saliency arm (89.7%, *p* < .05), Non-skippable microtutorial arm (89.4%, *p* < .05) and Skippable microtutorial arm (89.9%, *p* < .05) compared to the Control arm. There was no significant difference between the Default and Control arms (see Figure 16). It’s worth noting that in our sensitivity analysis, when analysing the same data but using an ordinal logistic regression instead of a binary logistic regression, there were no significant differences between the treatment arms and the Control arm (all *p* > .05). For the ordinal breakdown by treatment arm, see Figure 17. A Brant test showed that the proportional odds assumption still holds (X^2 (30) = 32.2, *p* = .36).

The discrepancy between the binary analysis and the ordinal sensitivity analysis may be because of the small number of participants who selected “Not at all” (see figure 17). The frequency distribution across categories in the outcome is resultantly very unbalanced, whereas when the outcome is binarised this effect is reduced as “Not at all” and “A little” are reduced. This would have affected the power of the two models differently. An alternative explanation is that, despite our proportional odds test holding, there is still a different, stronger, association between the treatment arms and the outcome on the cut-off between “A little” and “Moderately”. In this instance, we would suggest taking our main - binary - analysis as leading in the interpretation. However, as with all exploratory analyses, we advise caution in the interpretation and to use this for generation of new research hypotheses rather than as a solid foundation for policy recommendations.

Figure 16. The results of a logistic regression of exploratory outcome 3, comparing the percentage of participants who found the content settings moderately or very easy to understand in the Control arm to each treatment arm.



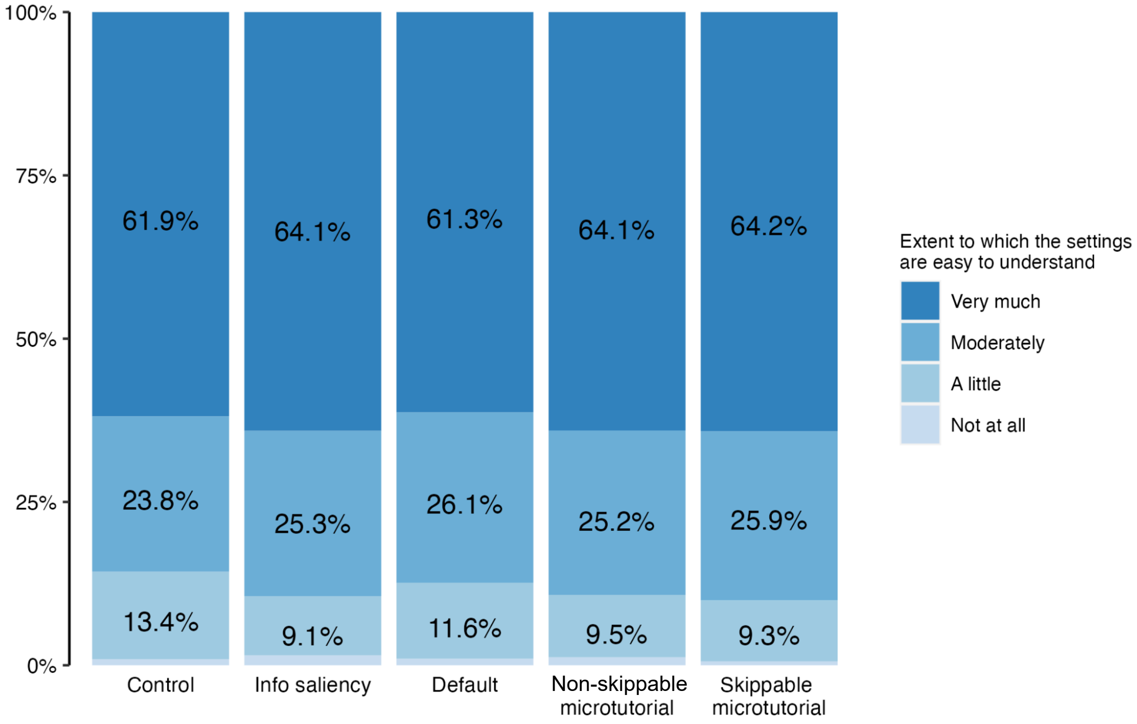
** p < .01, * p < .05, + p < .1

Significance and confidence intervals (95%) are not corrected for multiple comparisons.

Treatment bars show the mean of the control + the coefficient of the respective treatment variable.

Regression controls for age, gender, income, ethnicity and platform use.

Figure 17. The results of an ordinal regression of exploratory outcome 3, comparing the results of how easy to understand the participants said the content settings were in the Control arm to each treatment arm.



** p < .01, * p < .05, + p < .1

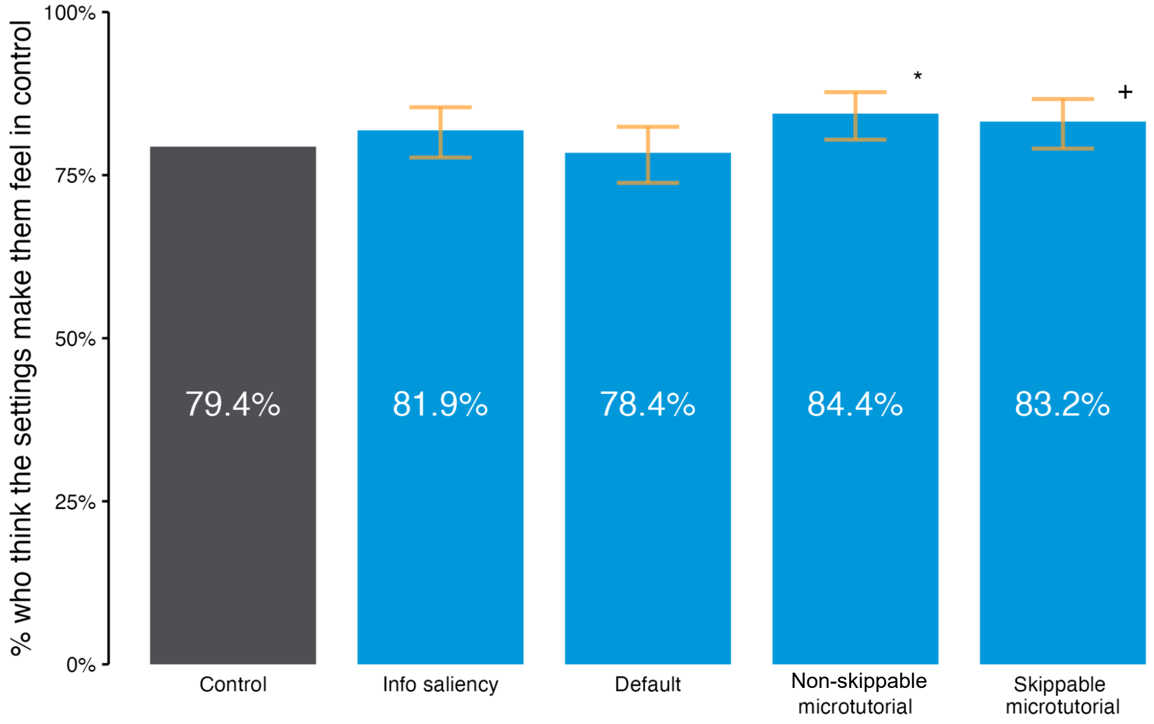
Significance is not corrected for multiple comparisons.

Regression controls for age, gender, income, ethnicity and platform use.

Feeling of control

81% said they moderately or very much think the content settings made them feel in control. This was significantly higher for participants in the Non-skippable microtutorial arm (84.4%, *p* < .05) compared to the Control arm (79.4%). There were no significant differences between the Control arm and other treatment arms, *p* < .05. Results are shown in Figure 18.

Figure 18. The results of a logistic regression of exploratory outcome 4, comparing the percentage of participants who said the content settings made them feel moderately or very much in control in the Control arm to each treatment arm.



** p < .01, * p < .05, + p < .1

Significance and confidence intervals (95%) are not corrected for multiple comparisons.

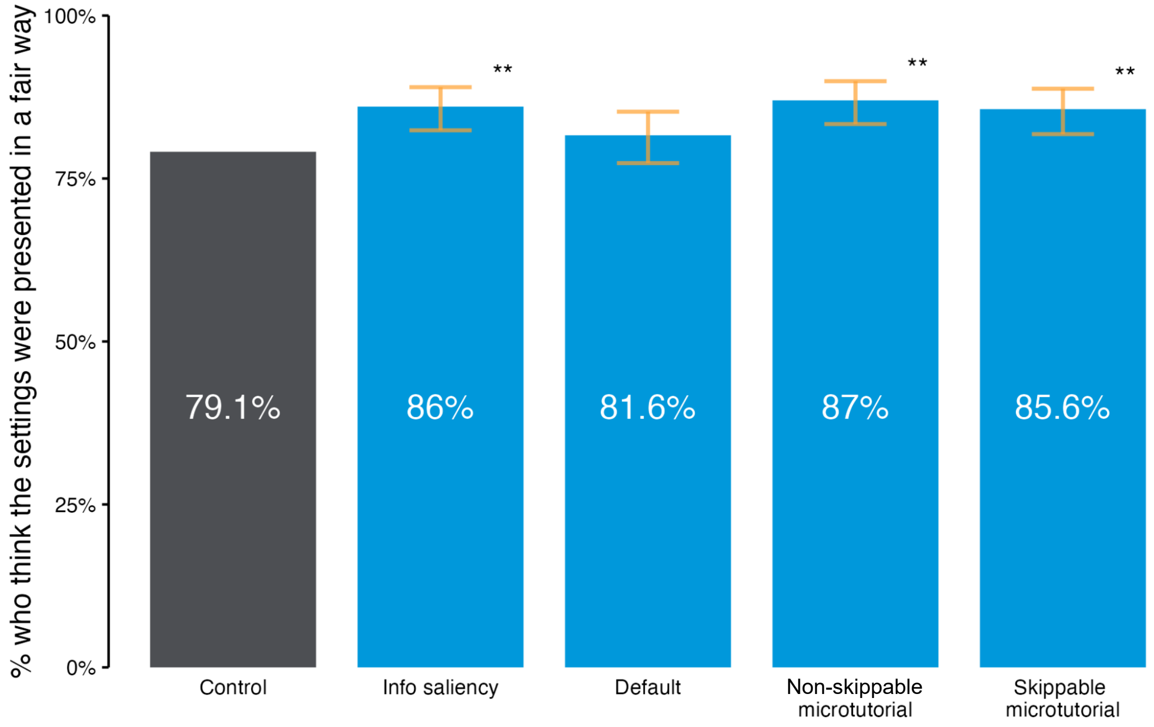
Treatment bars show the mean of the control + the coefficient of the respective treatment variable.

Regression controls for age, gender, income, ethnicity and platform use.

Presented in a fair way

83% said they moderately or very much think the content settings were presented in a fair way. This was significantly higher in the Info saliency (86%, *p* < .01), Non-skippable microtutorial (87%, *p* < .01) and Skippable microtutorial (85.6%, *p* < .01) arms that in the Control arm (79.1%). There was no significant difference between the Default arm and the Control arm. Results are shown in Figure 19.

Figure 19. The results of a logistic regression of exploratory outcome 5, comparing the percentage of participants who moderately or very much think the content settings are presented in a fair way in the Control arm to each treatment arm.



** p < .01, * p < .05, + p < .1

Significance and confidence intervals (95%) are not corrected for multiple comparisons.

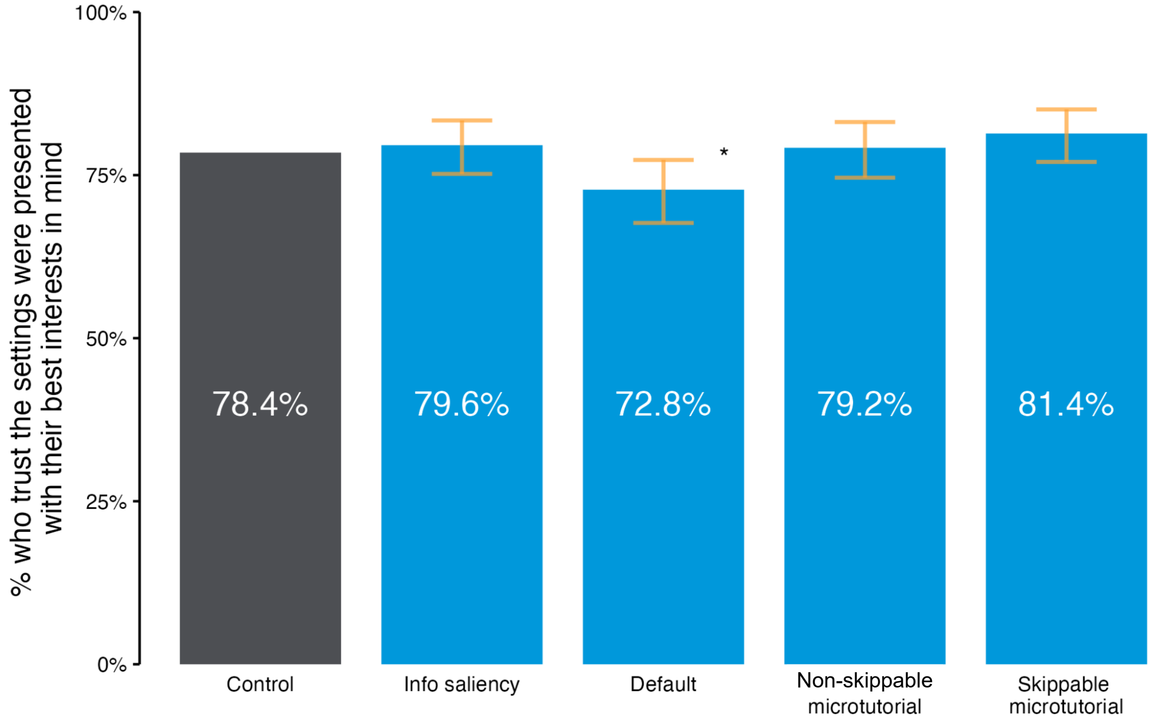
Treatment bars show the mean of the control + the coefficient of the respective treatment variable.

Regression controls for age, gender, income, ethnicity and platform use.

Trust they were presented with their best interests in mind

78% said they moderately or very much think the content settings were presented with their best interests in mind. This was significantly lower for those in the Default arm (72.8%, *p* < .05) than those in the Control arm (78.4%). There was no significant difference between the Control and the other treatment arms. Results are shown in Figure 20.

Figure 20. The results of a logistic regression of exploratory outcome 6, comparing the percentage of participants who moderately or very much trust that the content settings are presented with their best interests in mind in the Control arm to each treatment arm.



** p < .01, * p < .05, + p < .1

Significance and confidence intervals (95%) are not corrected for multiple comparisons.

Treatment bars show the mean of the control + the coefficient of the respective treatment variable.

Regression controls for age, gender, income, ethnicity and platform use.

4.5 Exploratory descriptives

4.5.1 Behaviour on the platform and controls screen

Participants spent a median of 2 minutes and 45 seconds on the WeConnect platform, including the sign-up and feed. They spent a median of 1 minute and 35 seconds on the WeConnect feed. During the sign-up, participants spent a median of 6 seconds choosing their content controls (7 seconds in the Control arm, 9 seconds in the Info saliency arm, 6 seconds in the Default arm, 6 seconds in the Non-skippable microtutorial arm and 6 seconds in the Skippable microtutorial arm[17]). Across arms, only 20 participants clicked to learn more about sensitive content when choosing their content controls (5 in the Control arm, 6 in the Default arm, 4 in the Non-skippable microtutorial arm, 5 in the Skippable microtutorial arm; there was no option to click to learn more in the Info saliency arm since the information was presented up front).

[17] Time spent in the microtutorials is excluded from the times of these arms.

Sentiment about WeConnect was positive: 61% of participants said WeConnect felt moderately or very similar to platforms they'd used before and 90% found WeConnect moderately or very easy to use.

4.5.2 Decision to change

In the follow-up survey, participants were asked why they chose to change or continue with the content settings they originally chose after browsing through WeConnect.

The top reasons participants chose to continue with the original choice (*n* = 3,069) was that they thought it was the right option for them (48%), the content they saw matched their expectations of the choice they originally made (34%) and that they liked the content they saw (26%). Only 9% said they did not understand how changing their choice would change the content they saw. The full list of response results is in Table 3.

Table 3. Why participants say they chose to continue with their original choice.

Why did you choose to continue seeing [all content types/reduced sensitive content] on WeConnect? (Participants could select more than one option, n = 3,069)	
I think it was the right option for me	48%
The content I saw matched my expectations of the choice I originally made	34%
I liked the content I saw on WeConnect	26%
I was worried I would be missing out on content that I'd like to see	18%
I don't care about the content I see on WeConnect	14%
I don't understand how changing my choice would change the content I see	9%
Other	3%

Other reasons participants gave for continuing with their content settings included that they believe in freedom of speech, that it's better to be aware of what's out there and that you can still scroll past content you don't want to see ("Whilst I disagree with just about everything I saw on there, I still believe in freedom of speech", "Seeing all available content offers a wider view of the world and the people in it – whether we agree with their views or not I feel it is better to be aware of them than not", "With any of the social media platforms there is no guarantees of pleasing everyone, so my choice was if you see it all then 'it is what it is' and if you make the choices to scroll on or not through content you may or may not like then the choice is still available in whether you want to be offended or not."). Some participants who had chosen to see reduced sensitive content also said that since they saw some sensitive content, they wouldn't want to see anymore by changing their settings ("Given I still saw unsuitable content with the filter on, I didn't want to see anything worse by taking it off").

On the other hand, participants who chose to change their settings (*n* = 431) say they did so because they were curious to see what would change (38%), they saw content that upset them (32%) or that they did not like the content on WeConnect (30%). Only 7% said they did not understand the choices when they initially made them. The full list of answers are in Table 4 (although note these are from a small sample size; *n* = 431).

Table 4. Why participants say they chose to change their settings from their original choice.

Why did you choose to change your settings to see [all content types/ reduced sensitive content] on WeConnect? (Participants could select more than one option, *n* = 431)	
I was curious to see what would change	38%
I saw content that upset me	32%
I didn’t like the content on WeConnect	30%
The content I saw didn’t match my expectations of the original choice	21%
I was worried I was missing out on content that I’d like to see	15%
I didn’t understand the choices when I initially made them	7%
I wasn’t paying attention when I made my original choice	3%
Other	5%

Other reasons participants gave for changing their content settings was that they content they saw was too much for them (“I don't mind seeing sensitive content but too much gets me mad”, “The content I saw without the restrictions wasn't what I expected and decided I didn't want to engage with it”). Some mentioned that they would have reported it instead if they had the option (“I saw some racist, ignorant comments that I would have reported if there was an option to do that”).

The decision to continue with or change their choice was not different between participants who originally chose to see all content types or reduced sensitive content (88% and 87% respectively).

4.5.3 Skipping the microtutorial

Of those in the Skippable microtutorial arm (*n* = 664), 73% skipped the tutorial. Most participants did so in the first two screens (29% in the arm skipped on the first screen and 26% on the second screen).

Participants in this arm were asked why they chose to skip or follow the microtutorial. For those who skipped the microtutorial (n = 484), the biggest reasons for doing so were that they did not think they needed a tutorial (58%) and that they already know about sensitive content (43%). The full list of answers is in Table 5.

Table 5. Why participants skipped the microtutorial.

You chose to skip the tutorial about sensitive content and the available choice options. What were the main reasons? (Participants could select more than one option, *n* = 484)	
I just didn't think I needed a tutorial	58%
I already knew about sensitive content	43%
It was too simple	13%
I got bored	9%
It was too long	7%
It was too slow	6%
It was poorly designed	3%
Other (e.g. "I felt that I knew what sensitive content would be.")	1%

For those who did not skip the microtutorial (n = 180), 55% said they did so because they wanted to learn more about sensitive content, 44% said it was easy to follow and 32% said they thought it was required for this study. Only 8% said they wanted to skip it but did not see how. The full list of answers are in Table 6.

Table 6. Why participants did not skip the microtutorial.

You chose to follow the tutorial about sensitive content and the available choice options. What were the main reasons? (Participants could select more than one option, *n* = 180)	
I wanted to learn more about sensitive content	52%
It was easy to follow	44%
I thought it was required for this study	32%
I found it engaging	25%
I liked the design	15%
I wanted to skip it but didn't see how	8%
Other (e.g. "I wanted to see what the platform is saying about it.")	< 1%

4.5.4 Sentiment to Non-skippable microtutorial

Participants in the Non-skippable microtutorial arm (*n* = 651) were asked about their sentiment towards the tutorial. Generally, they had positive opinions on the microtutorial: 66% said it was easy to understand and 34% said it helped them to learn more about sensitive content. 9% said they wish they could have skipped the microtutorial. The full list of answers are in Table 7.

Table 7. How do participants describe the experience of the Non-skippable microtutorial.

You followed the tutorial about sensitive content and the available choice options. Which of the following best describes your experience? (Participants could select more than one option, *n* = 651)	
It was easy to follow	66%
It helped me learn more about sensitive content	34%
I found it engaging	23%
I liked the design	22%
I wish I could skip it	9%
It was boring	6%
It was too long	5%
It was annoying	4%
Other (e.g. “I don’t remember it.”)	2%

4.5.5 Comprehension

When looking at the individual pieces of content analysed in secondary analysis 1, participants were better at identifying violence and hate speech as sensitive content than they were misinformation and misogyny. 80% correctly said a violent video and 76% correctly said that a short text post including hate speech were sensitive content. 58% correctly said a link to an inaccurate blog post and 52% correctly said a video expressing misogynistic views was sensitive content.

Generally, participants were better at correctly identifying non-sensitive content. 90% said a photo of someone waving a flag, 85% said a photo of a dog chasing a cat and 83% said a news article about the outcome of a lawsuit about election fraud was not sensitive content. Participants were less accurate when categorising a video of a drunk user chugging three beers in a row; 65% correctly said this was not sensitive content.

4.5.6 Previous experience with content controls

Participants who used social media (*n* = 3,180) were also asked questions to understand their previous experience with content controls. 26% said they had changed the settings that determine what kind of content they see.

The biggest barriers to changing content settings were related to a lack of interest. 31% say they were happy with the content they currently saw and 18% say they did not think they needed content controls. The full list of responses is in Table 8.

Table 8. Barriers participants have experienced that have prevented them from changing the settings that determine what kind of content they see on social media.

What, if anything, has prevented you from changing the settings that determine what kind of content you see on social media? (Participants could select more than one option, *n* = 3,180)	
I am happy with the content I currently see	31%
I don’t think I need any content controls	18%
I didn’t know I could change the settings	13%
I don’t want to	12%
I never get around to changing the settings	11%
Not to miss out on the content I want to see	11%
I don’t know how to	9%
I don’t trust how the platform categorises content	8%
It’s hard to change the settings	5%
It doesn’t do anything	5%
I don’t understand how this would change my feed	4%
Other (e.g. “I block certain users and mute certain topics as and when I find across them.”, “I only don’t want to see specific things so i just block hashtags or words instead of general sensitive content”, “Only really use it to keep in touch with family”)	1%
Nothing has prevented me from changing the settings on what kind of content I see (exclusive)	19%

19% said they’ve never experienced anything that has prevented them from changing the settings on what kind of content they see. This was higher for those who have changed their settings before (32%, *n* = 854) than those who have not (14%, *n* = 2,160). More participants who have changed their settings before said not trusting how platforms categorise content was a barrier to changing their settings (14% compared to 7% of those who have never changed their settings). Those who have never changed their settings before were more likely to say that they did not think they needed to (22% compared to 9% of those who have changed their settings), and that they did not want to (15% compared to 4%). A full breakdown of barriers by these subgroups is in Table 9.

Table 9. Barriers participants have experienced that have prevented them from changing the settings that determine what kind of content they see on social media by whether participants have changed their settings before or not (excluding 166 who said they didn't know if they'd changed their settings).

	Have changed their content settings before (*n* = 854)	**Have never changed their content settings (*n* = 2,160)**
I am happy with the content I currently see	25%	33%
I don't think I need any content controls	9%	22%
I didn't know I could change the settings	5%	15%
I don't want to	4%	15%
I never get around to changing the settings	6%	13%
Not to miss out on the content I want to see	12%	11%
I don't know how to	4%	10%
I don't trust how the platform categorises content	14%	7%
It's hard to change the settings	8%	4%
It doesn't do anything	7%	4%
I don't understand how this would change my feed	5%	4%
Nothing has prevented me from changing the settings on what kind of content I see (exclusive)	32%	14%

5. Summary and Limitations

5.1 Summary

The focus of the trial was to see if any of the interventions would affect the alignment of content choices made by participants with their underlying preferences. We sought to measure this through whether participants stuck to their original choice or changed their initial content settings. If participants changed their content settings, then this could suggest that they were unhappy with their original content choice.

Across all trial arms, roughly 88% of participants chose to continue with the choice they made at the sign-up stage of the experiment. There were no significant differences between the intervention arms and the Control. The likelihood of a participant continuing with their original choice was the same for participants who chose to see “All content types” and “Reduced sensitive content”. This suggests that none of the interventions was more effective than the Control at increasing the likelihood of participants sticking to their initial content settings.

Nevertheless, our exploratory analysis found differences in the initial choice of settings at the sign-up stage. Participants in the default arm were significantly less likely to choose “Reduced sensitive content” compared to the Control group. While the sample size for this analysis was low, the reasons given in the default group for continuing or changing the settings did not substantially differ from the reasons given in the other arms. This finding suggests that participants’ initial choice is influenced by the choice architecture deployed, but even then, participants generally stick to the original content choices they make – regardless of what that choice was and how it was presented.

Similarly, the main reason provided by participants who continued with their initial choice in our experiment was that they still believed it to be the right option for them. The top reason for changing the setting was curiosity about what would change rather than a mismatch of the content they saw with their initial expectations.

Comprehension of what counts as sensitive content did not differ across arms. In general, participants did well at discerning sensitive and non-sensitive pieces of content in our comprehension task. Participants did worse at correctly classifying misinformation and misogynistic content as sensitive, which could be explained by subjective perceptions of the boundaries of freedom of speech, especially given that very few participants clicked through to learn more about what comprises sensitive content on this platform.

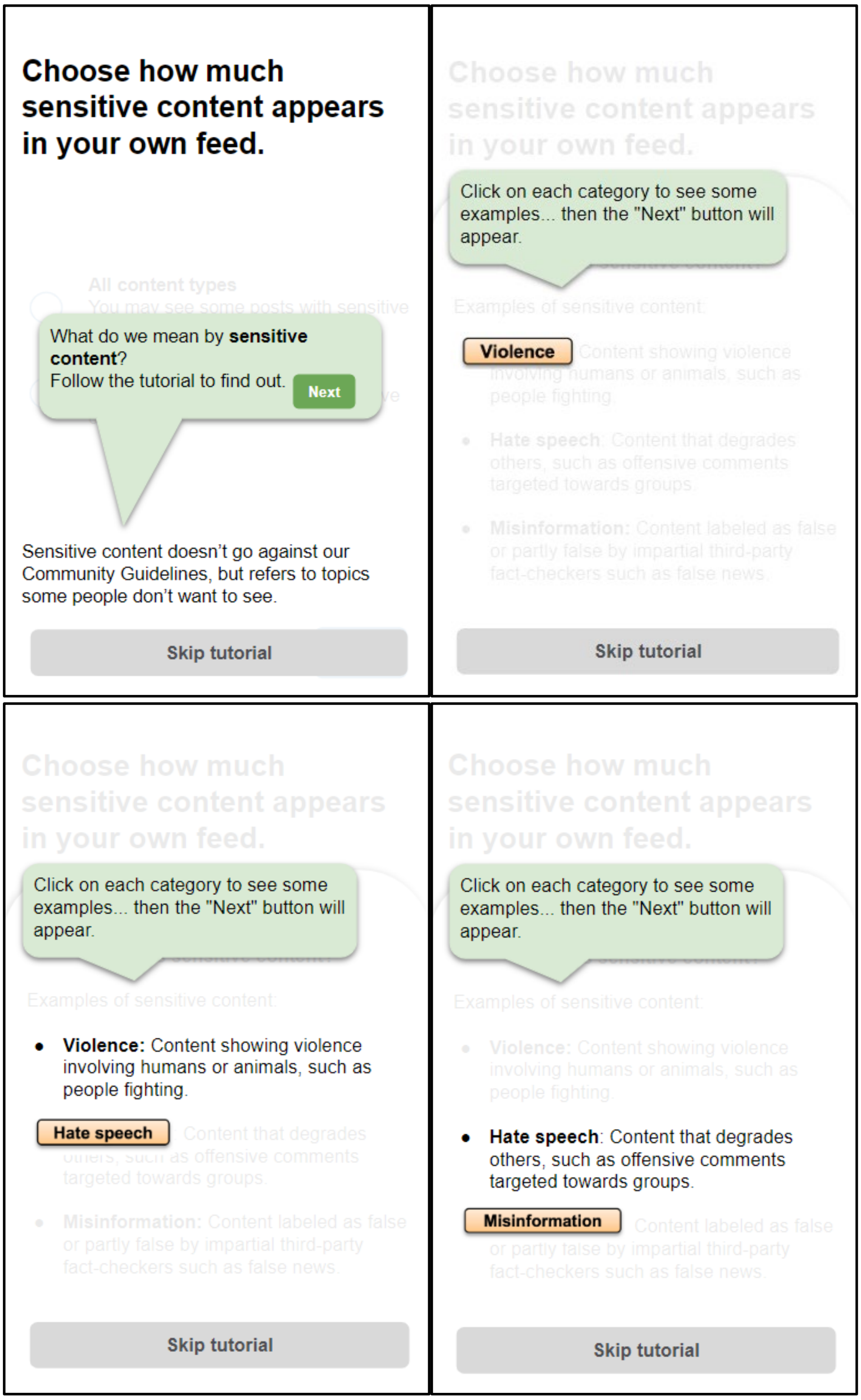
Overall sentiment scores towards the way the sensitive control settings were presented did not differ between conditions. However, our exploratory analysis found that more participants in the Info saliency and Microtutorial conditions found the settings easier to understand and presented in a fair way compared to the Control group. Further, participants in the Microtutorial arms felt more in control over their content settings than the participants in the Control arm. Participants in the Default arm were less likely to perceive the content settings to have been presented to them with their best interest in mind than Control group participants.

5.2 Limitations

Given the environment we ran our experiment in, several limitations apply to our findings. No matter how carefully designed, a simulated platform is not able to fully replicate the incentives and motivations that guide users' behaviours on social media. Importantly, real-world sensitive content may include content that is more harmful and more personalised than the content shown in our research (see section 3.4 for further details on content selection). Moreover, the short timescale at which our online experiment had to measure outcomes limits the conclusions that can be drawn with respect to the long-term effects of our interventions. Despite these limitations, we believe online RCTs are a useful tool for building the evidence base on what works to increase online safety.

6. Annex

Annex A: Interactive microtutorial



Choose how much sensitive content appears in your own feed.

All content types
You may see some posts with sensitive

What do we mean by **sensitive content**?
Follow the tutorial to find out.

Sensitive content doesn't go against our Community Guidelines, but refers to topics some people don't want to see.

Choose how much sensitive content appears in your own feed.

Click on each category to see some examples... then the "Next" button will appear.

Examples of sensitive content:

Content showing violence involving humans or animals, such as people fighting.

- **Hate speech**: Content that degrades others, such as offensive comments targeted towards groups.
- **Misinformation:** Content labeled as false or partly false by impartial third-party fact-checkers such as false news.

Choose how much sensitive content appears in your own feed.

Click on each category to see some examples... then the "Next" button will appear.

Examples of sensitive content:

- **Violence:** Content showing violence involving humans or animals, such as people fighting.

Content that degrades others, such as offensive comments targeted towards groups.

- **Misinformation:** Content labeled as false or partly false by impartial third-party fact-checkers such as false news.

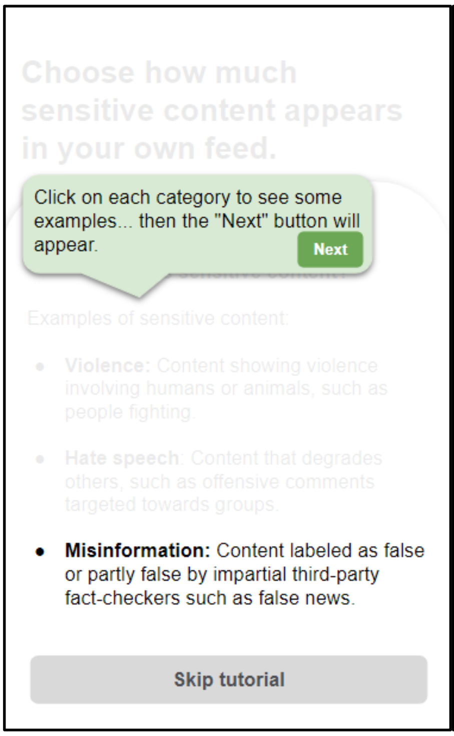
Choose how much sensitive content appears in your own feed.

Click on each category to see some examples... then the "Next" button will appear.

Examples of sensitive content:

- **Violence:** Content showing violence involving humans or animals, such as people fighting.
- **Hate speech**: Content that degrades others, such as offensive comments targeted towards groups.

Content labeled as false or partly false by impartial third-party fact-checkers such as false news.

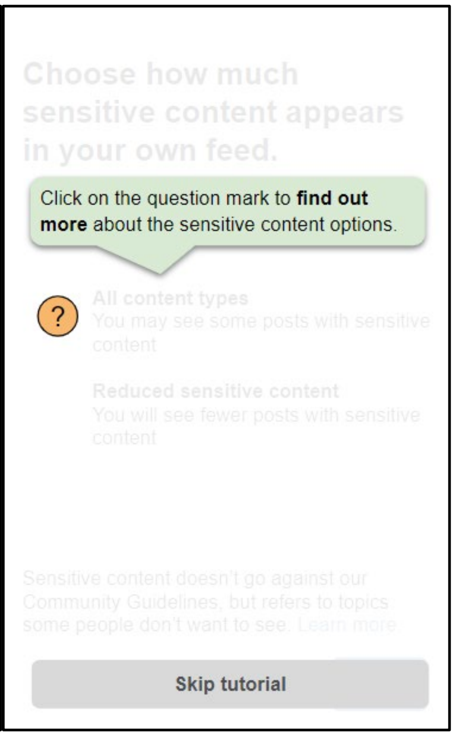


Choose how much sensitive content appears in your own feed.

Click on each category to see some examples... then the "Next" button will appear.

Examples of sensitive content:

- **Violence:** Content showing violence involving humans or animals, such as people fighting.
- **Hate speech**: Content that degrades others, such as offensive comments targeted towards groups.
- **Misinformation:** Content labeled as false or partly false by impartial third-party fact-checkers such as false news.



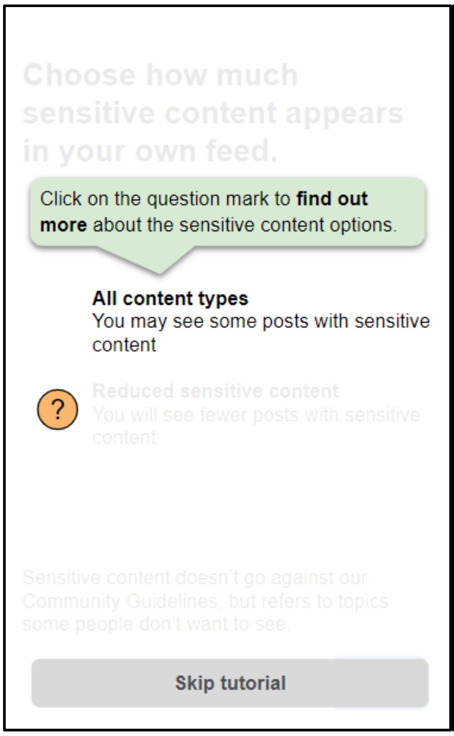
Choose how much sensitive content appears in your own feed.

Click on the question mark to **find out more** about the sensitive content options.

All content types
You may see some posts with sensitive content

Reduced sensitive content
You will see fewer posts with sensitive content

Sensitive content doesn't go against our Community Guidelines, but refers to topics some people don't want to see. Learn more



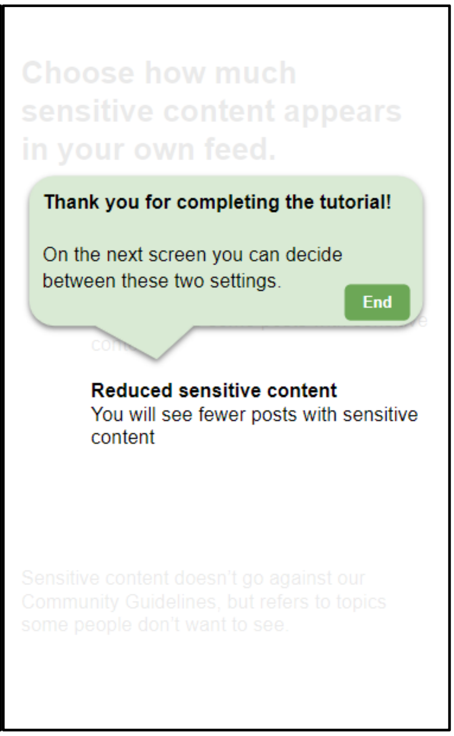
Choose how much sensitive content appears in your own feed.

Click on the question mark to **find out more** about the sensitive content options.

All content types
You may see some posts with sensitive content

Reduced sensitive content
You will see fewer posts with sensitive content

Sensitive content doesn't go against our Community Guidelines, but refers to topics some people don't want to see.



Choose how much sensitive content appears in your own feed.

Thank you for completing the tutorial!

On the next screen you can decide between these two settings.

Reduced sensitive content
You will see fewer posts with sensitive content

Sensitive content doesn't go against our Community Guidelines, but refers to topics some people don't want to see.

Annex B: Ordinal models

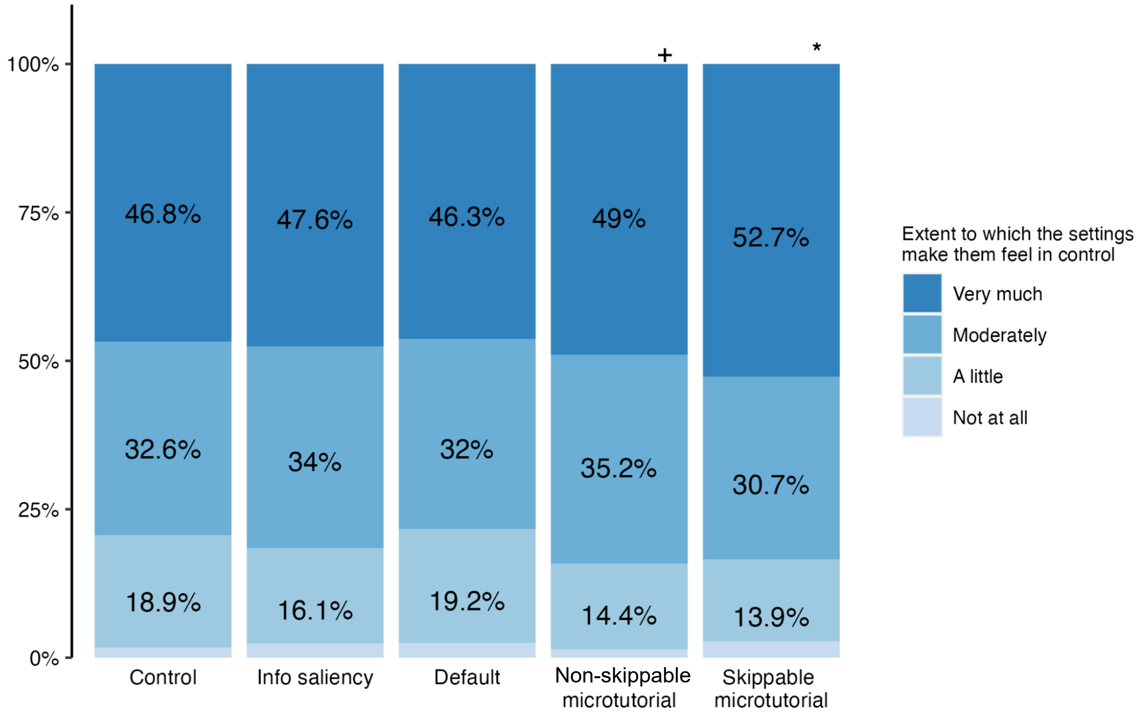
The Brant-Wald test assesses whether the assumption of proportional odds in an ordinal logistic regression model is valid by checking if the relationship between each predictor and the response is consistent across different levels of the response. A non-significant omnibus test suggests the assumption of proportional odds holds. The Brant-Wald test showed the proportional odds assumption generally held for ordinal regression models on exploratory outcomes 3 and 4 but not for exploratory outcomes 5 and 6 (see Table 10). Results from these regressions should be interpreted with caution.

Table 10. The results of the Brant test on the proportional odds assumption for each sentiment variable.

Exploratory outcome 3: Ease of understanding			
	χ^2	**df**	***p***
Omnibus	32.2	30	.36
Info saliency	6.11	2	.05
Default	1.58	2	.45
Non-skippable microtutorial	3.99	2	.14
Skippable microtutorial	4.23	2	.12
Exploratory outcome 4: Feeling of control			
	χ^2	**df**	***p***
Omnibus	31.94	30	.37
Info saliency	2.10	2	.35
Default	0.84	2	.66
Non-skippable microtutorial	3.10	2	.21
Skippable microtutorial	4.21	2	.12
Exploratory outcome 5: Presented in a fair way			
	χ^2	**df**	***p***
Omnibus	48.55	30	.02
Info saliency	4.11	2	.13
Default	3.04	2	.22
Non-skippable microtutorial	12.66	2	.00
Skippable microtutorial	6.65	2	.04
Exploratory outcome 6: Trust they were presented with their best interests in mind			
	χ^2	**df**	***p***

Omnibus	46.42	30	.03
Info saliency	0.10	2	.95
Default	4.06	2	.13
Non-skippable microtutorial	0.21	2	.90
Skippable microtutorial	1.80	2	.41

Figure 21. The results of an ordinal regression of exploratory outcome 4, comparing the results of how much the settings make them feel in control in the Control arm to each treatment arm.

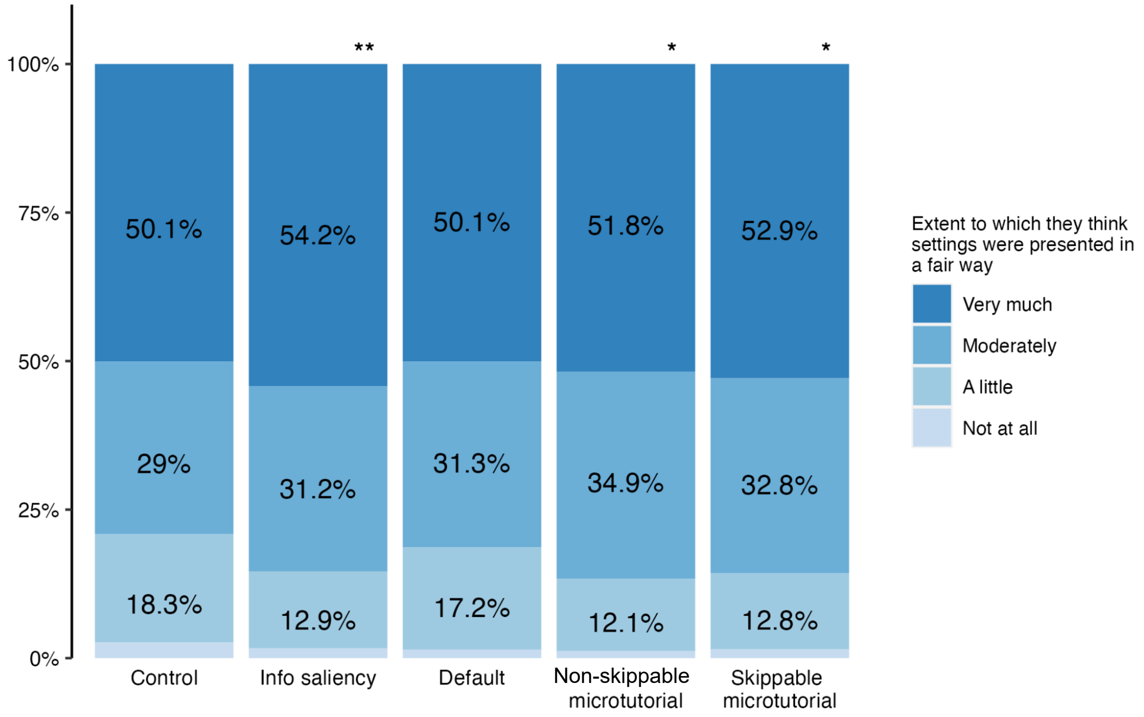


** p < .01, * p < 0.5, + p < .1

Significance is not corrected for multiple comparisons.

Regression controls for age, gender, income, ethnicity and platform use.

Figure 22. The results of an ordinal regression of exploratory outcome 5, comparing the results of how much participants think the settings were presented in a fair way in the Control arm to each treatment arm.

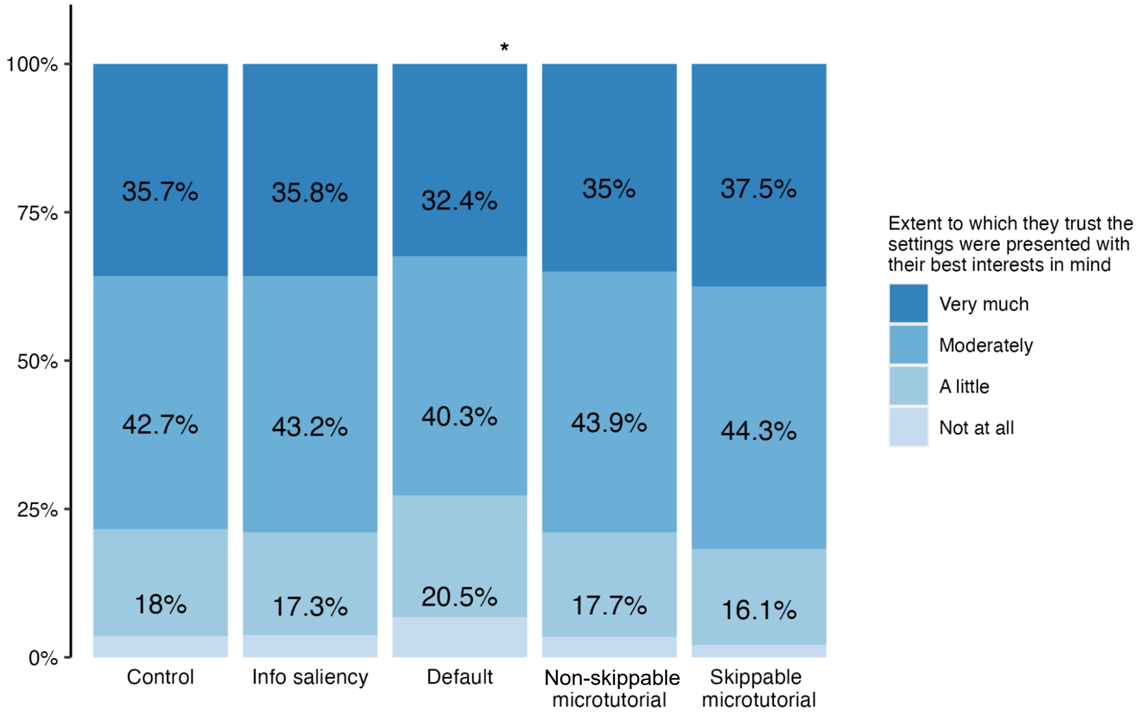


** p < .01, * p < 0.5, + p < .1

Significance is not corrected for multiple comparisons.

Regression controls for age, gender, income, ethnicity and platform use.

Figure 23. The results of an ordinal regression of exploratory outcome 5, comparing the results of how much participants trust the settings were presented with their best interests in mind in the Control arm to each treatment arm.



** p < .01, * p < 0.5, + p < .1

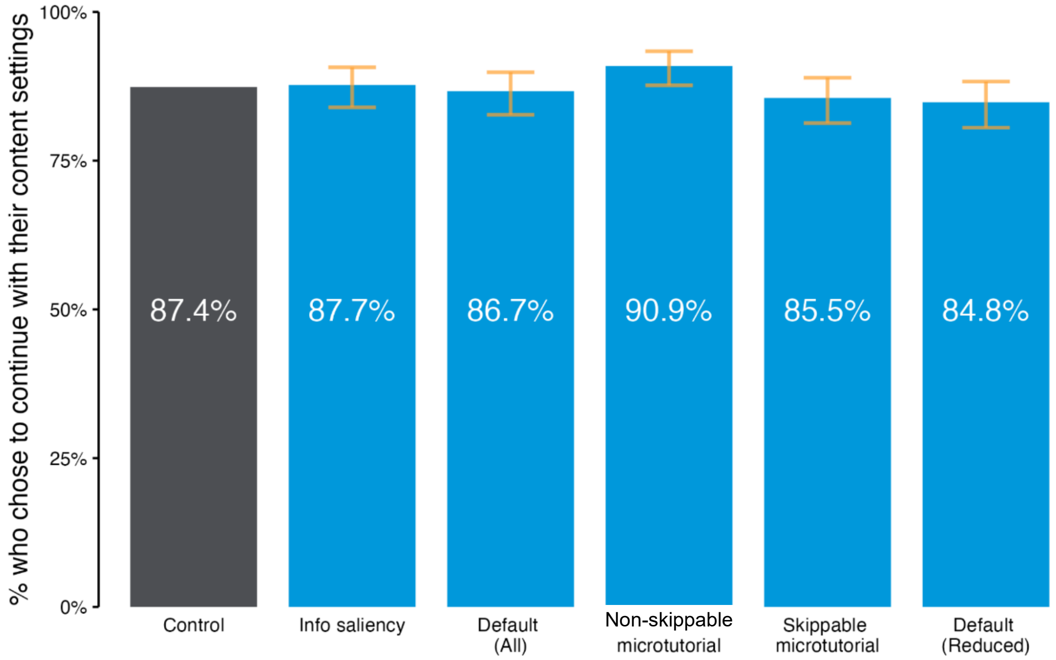
Significance is not corrected for multiple comparisons.

Regression controls for age, gender, income, ethnicity and platform use.

Annex C: Additional exploratory mini-experiment

Following the completion of the trial, BIT ran an exploratory mini-experiment to investigate the effect of pre-selecting a different option on the initial choice page. An additional 700 participants were recruited, and all were allocated to have “Reduced sensitive content” pre-selected on the initial choice page. These findings are exploratory, and analysis was not corrected for multiple comparisons.

Figure 24. The results of additional exploratory analysis, comparing the percentage of participants who chose to continue with their content settings in the Control arm to each intervention arm, with an extra addition of Default (Reduced) arm.



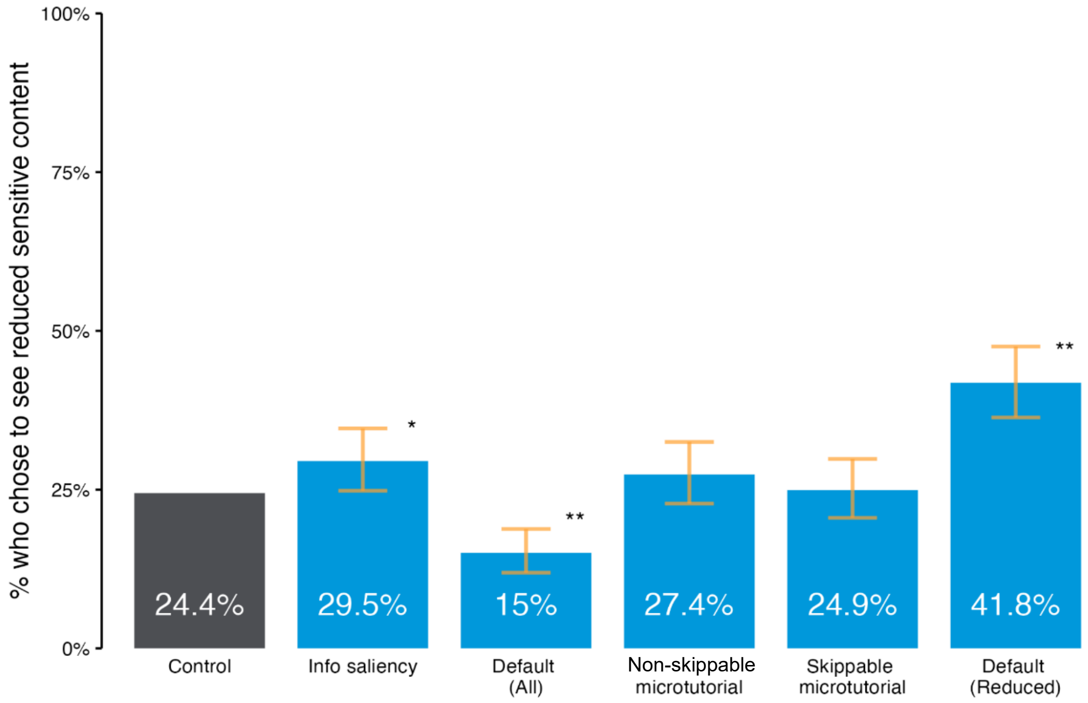
** p < .01, * p < .05, + p < .1

Significance is corrected for multiple comparisons. Confidence intervals (95%) are not corrected for multiple comparisons.

Regression controls for age, gender, income, ethnicity and platform use.

Treatment bars show the mean of the control + the coefficient of the respective treatment variable.

Figure 25. ***The results of additional exploratory analysis, comparing the percentage of participants who chose to see reduced sensitive content in the Control arm to each intervention arm,*** **with an extra addition of Default (Reduced) arm.**



** p < .01, * p < 0.5, + p < .1

Significance and confidence intervals (95%) are not corrected for multiple comparisons.

Treatment bars show the mean of the control + the coefficient of the respective treatment variable.

Regression controls for age, gender, income, ethnicity and platform use.