On-platform interventions for content controls

Technical Report with methodology and results

Prepared by the Behavioural Insights Team and Ofcom

19 April 2024

THE BEHAVIOURAL INSIGHTS TEAM

Contents

1. Background 4
1.1 Policy and regulatory context 4
1.1.1 Making Sense of Media 4
1.1.2 Regulatory context 4
1.2 Research objectives 4
2. Interventions and hypotheses 5
2.1 Trial arms overview 5
2.2 Control arm 6
2.3 Intervention arms 7
2.3.1 Timing 8
2.3.2 Messages 9
3. Methodology 11
3.1 Trial design 11
3.2 Simulated social media platform 11
3.2.1 Platform design and functionality 11
3.2.2 Training task 12
3.2.3 Main task 12
3.2.4 Stimuli 12
3.2.5 Post-feed survey 13
3.2.6 User testing 13
3.3 Sampling and data collection 14
3.3.1 Sample criteria 14
3.3.2 Power calculations 14
3.3.3 Data collection 14
3.4 Ethical considerations 15
3.5 Analytical framework 17
3.5.1 Data checks 17
3.5.2 Analytical strategy 17
3.5.3 Primary analysis 17
3.5.4 Secondary analyses 18
3.5.5 Exploratory analyses 20
4. Results 23
4.1 Sample characteristics 23

4.2 Primary analysis: Whether people checked their content settings 24
4.3 Secondary analyses 26
4.3.1 Further comparisons on the primary outcome 26
4.3.2 Whether their settings match their preferences 26
4.4 Exploratory analyses 28
4.4.1 Recall 28
4.4.2 Final choice 30
4.4.3 How many times participants reviewed their settings 30
4.4.4 Time viewing prompt 31
4.4.5 Time reviewing 31
4.4.6 Sentiment 31
4.4.7 Primary outcome by psychological variables 34
4.4.8 Primary and secondary analysis excluding those who saw the prompt at the end of the feed 35
4.5 Exploratory Descriptives 36
4.5.1 Gear icon 36
4.5.2 Clickthroughs 36
4.5.3 Decision to check 36
4.5.4 Sentiment to the Control arm 37
4.5.5 Previous experience with content settings 37
5. Summary and Limitations 39
5.1 Summary 39
5.2 Limitations 40
6. Annex 41
Annex A: Ordinal models 41
Annex B: User journey 43

1. Background

1.1 Policy and regulatory context

1.1.1 Making Sense of Media

Making Sense of Media (MSOM) is Ofcom's programme of work to help improve the online skills, knowledge and understanding of UK adults and children.[1] The MSOM team achieves this by sharing evidence-based insights, encouraging the media literacy community to pilot activities and initiatives which support MSOM's aim. The MSOM programme has a dual focus on people and platforms. MSOM's work with platforms/online services aims to establish what works well online and what does not. Recently this has involved the development of Best Practice Principles for Media Literacy by Design that provide social media, gaming, pornography, sharing and search services of all sizes with guidance on how to develop on-platform interventions to promote media literacy.[2] This research was conducted to support this work.

1.1.2 Regulatory context

Ofcom has a statutory duty to promote media literacy and to carry out research into media literacy matters.[3] 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.[4,5] Ofcom also recently became the UK's online safety regulator following the Online Safety Act 2023 (OSA) becoming law.[6] Ofcom's media literacy work will make an important contribution implementing the OSA, in particular the changes to Ofcom's media literacy duties. However, this work should not be interpreted as a statement of our policy on any guidance or our codes of practice under the OSA or prejudge any further work to develop policy in relation to that OSA.

1.2 Research objectives

Together with Ofcom's Behavioural Insight Hub, the Behavioural Insights Team (BIT) conducted a randomised control trial (RCT) to build evidence on on-platform interventions. We explored prompts to encourage people to make an active choice about controls that determine how much sensitive content[7] they see. Such controls are referred to as content controls or content settings in this research.

The trial tested different prompts to encourage users to review their content controls when viewing social media feeds. We tested prompts with i) different messages and ii) different timing.

---

[1] Ofcom, n.d. Making Sense of Media.
[2] Ofcom, n.d. Establishing best practice media literacy design principles.
[3] UK Parliament, 2003. Communications Act 2003.
[4] Ofcom, n.d. Making Sense of Media.
[5] Ofcom, n.d. Video-sharing platform (VSP) regulation.
[6] UK Parliament, 2023. Online Safety Act 2023.
[7] 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.

The interventions developed for this trial were not aiming to steer people towards a particular choice (such as choosing to see reduced sensitive content). Instead, the trial focused on encouraging users to make an active, informed choice about their content controls. This was measured by whether participants clicked through to check their content controls.

The trial aimed to answer the following main research questions:

- **RQ1**: Do on-platform prompts encourage users to review content settings?
- **RQ2**: Does the effect vary depending on the timing of the prompt?
- **RQ3**: Does the effect of the prompt vary depending on the prompt message?

Our primary outcome measure was whether a participant clicks through to check their content settings.

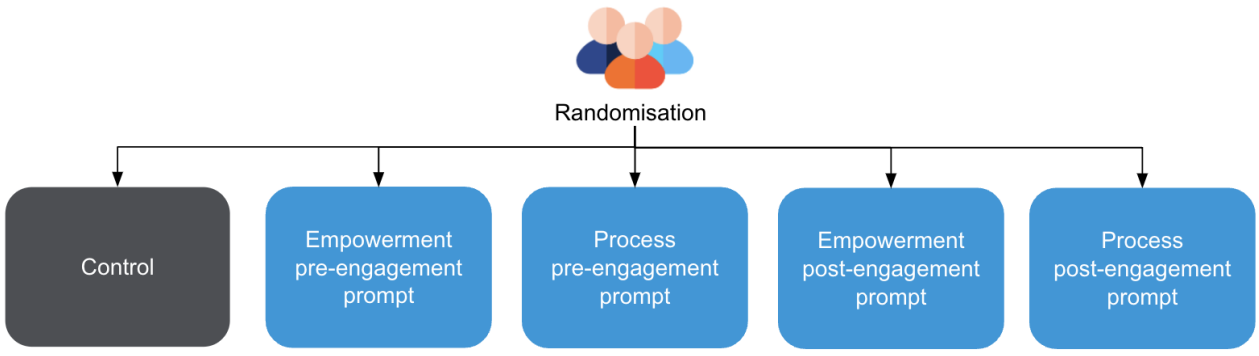
We also examined the impact of the prompts on the alignment between the content settings chosen by users in the experiment and their self-reported preferences for content settings outside of the experiment. We explored several other outcomes, such as the final choice of content settings and sentiment towards the prompts, to help us better understand the results of the primary and secondary analyses and generate hypotheses for future research (see section 3.5 for the full analytical framework).

2. Interventions and hypotheses

2.1 Trial arms overview

We conducted a five-arm trial with one control and four treatment[8] conditions. In this trial, we tested different messages and how effective they were at causing people to review their content controls. Figure 1 gives an overview of the trial arms into which participants were randomised.

Figure 1. Overview of trial arms.



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

2.2 Control arm

When designing the Control arm of the trial, we aimed to replicate a typical social media platform, with users able to review their content settings, at any point, through a gear icon on their feed (see Figure 2). No prompt was included in the Control arm. If users clicked on the gear icon, they were shown a pop-up which asked if they would like to check their content settings. They could access a definition of sensitive content by clicking ‘Learn more’ (see Figure 3).

Figure 2: WeConnect content feed.

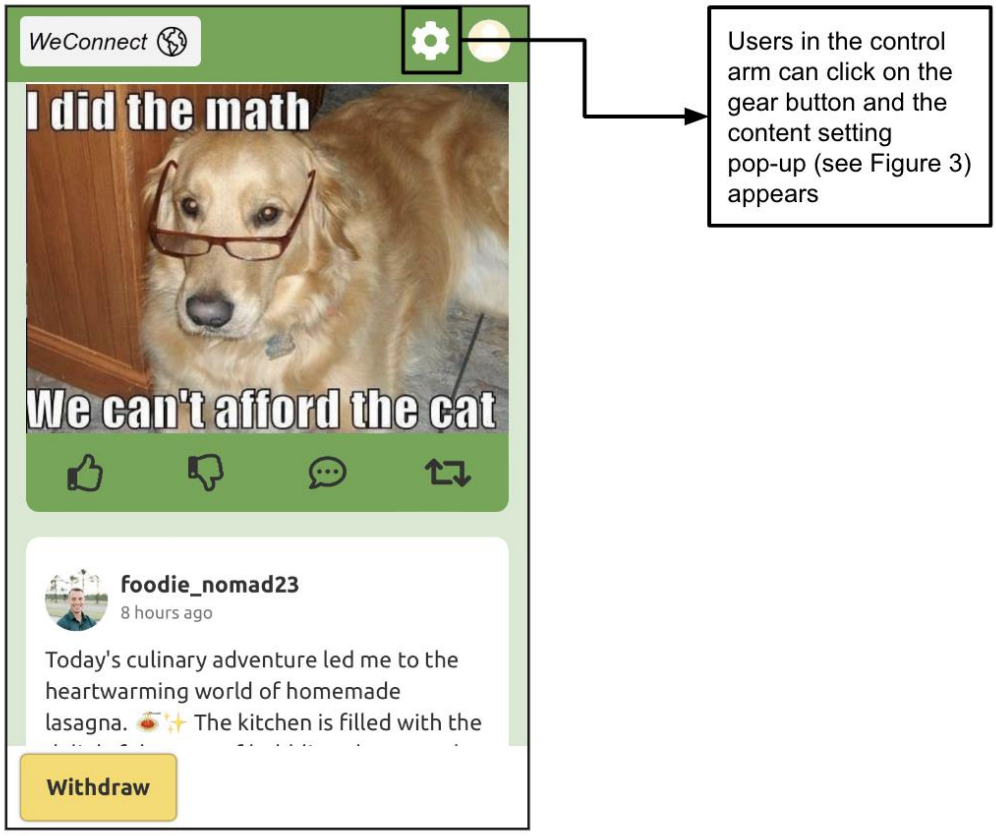
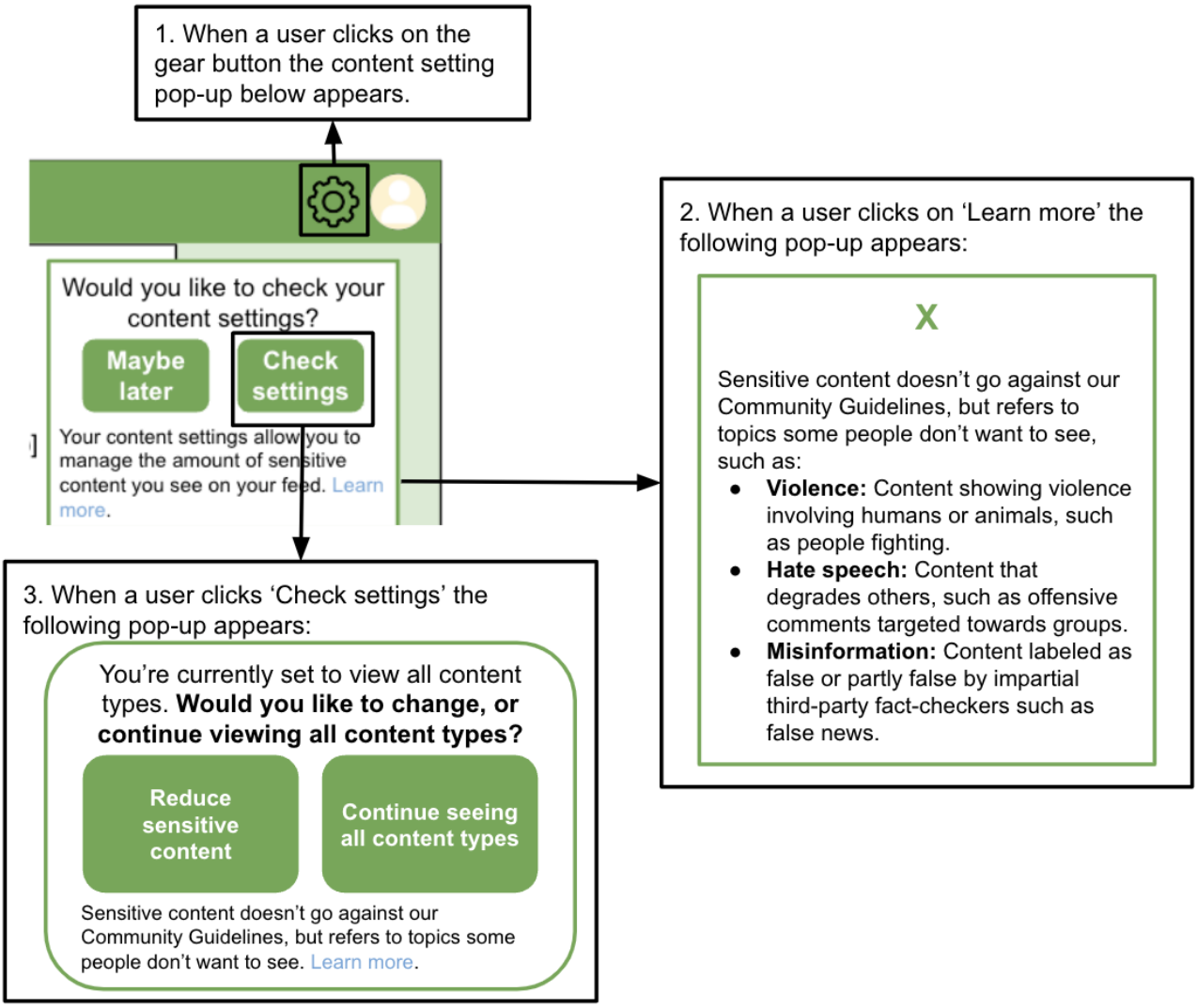


Figure 3: Content settings available when scrolling on simulated social media feed.



2.3 Intervention arms

The content of the message in the prompt as well as the timing of the prompt varied across intervention arms (see Table 1).

Table 1: Overview of message content and timing

	Empowerment message	Process message
Pre-engagement	**Arm 1: Pre-engagement & Empowerment** Prompt appears after the first post in their feed (the first post was always non-sensitive) with the following message: *"Your feed, your choice – you can choose the amount of sensitive content that you see."*	**Arm 2: Pre-engagement & Process** Prompt appears after the first post in their feed (the first post was always non-sensitive) with the following message: *"It takes just two steps to check and update your content settings."*
Post-engagement	**Arm 3: Post-engagement & Empowerment** Prompt appears after a participant dislikes a sensitive post or after scrolling through all sensitive posts in their feed with the following message: *"Your feed, your choice – you can choose the amount of sensitive content that you see."*	**Arm 4: Post-engagement & Process** Prompt appears after a participant dislikes a sensitive post or after scrolling through all sensitive posts in their feed with the following message: *"It takes just two steps to check and update your content settings."*

2.3.1 Timing

Prompts at timely moments when a user may be more receptive to a message can be an effective tool for behaviour change.[9] In the context of content controls, we theorised that these timely moments could be after a user has already **taken a small action** (in this case, disliking sensitive content) or/and when a user is in a particular **emotional state** (in this case, after having seen sensitive content).

Pre-engagement arms

In the Pre-engagement arms, participants received a prompt at the start of the simulated social media feed. The prompt was delivered after the first post, which was always non-sensitive. Our hypotheses are outlined below.

H1: The probability of reviewing content settings at least once in the Pre-engagement arms will be significantly higher compared to the Control.

Post-engagement arms

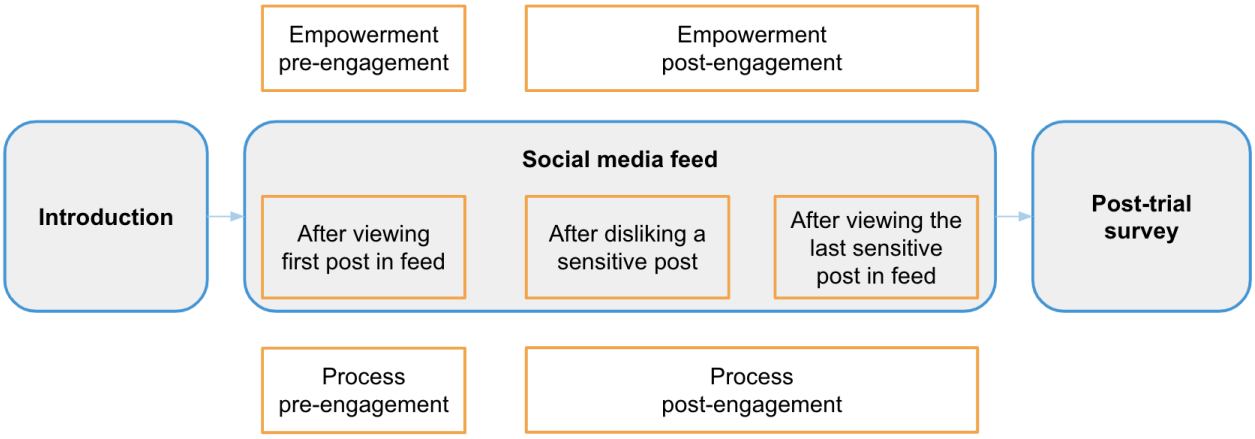
In the Post-engagement arms, participants received a prompt after they disliked a sensitive post or after they had viewed the last sensitive post in the feed.

H2: The probability of reviewing settings at least once in the Post-engagement arms will be significantly higher compared to the Control.

H3: The probability of reviewing settings at least once in the Post-engagement arms will be significantly higher compared to the Pre-engagement arms.

[9] BIT, 2015. EAST: Four simple ways to apply behavioural insights. [accessed 28 March 2024].

Figure 4: Stylised participant journey showing when Pre-engagement and Post-engagement messages are shown.



2.3.2 Messages

Decisions are influenced by how information is worded and what aspects are emphasised. The wording of the messages in this trial aimed to address potential motivational barriers. Participants saw either an Empowerment message or a Process message (see Figure 5).

Empowerment message arms

The Empowerment message focused on the emotional appeal and belief that a user can control their experience.[10,11]

H4: The probability of reviewing settings at least once in the Empowerment message arms will be significantly higher compared to the Control.

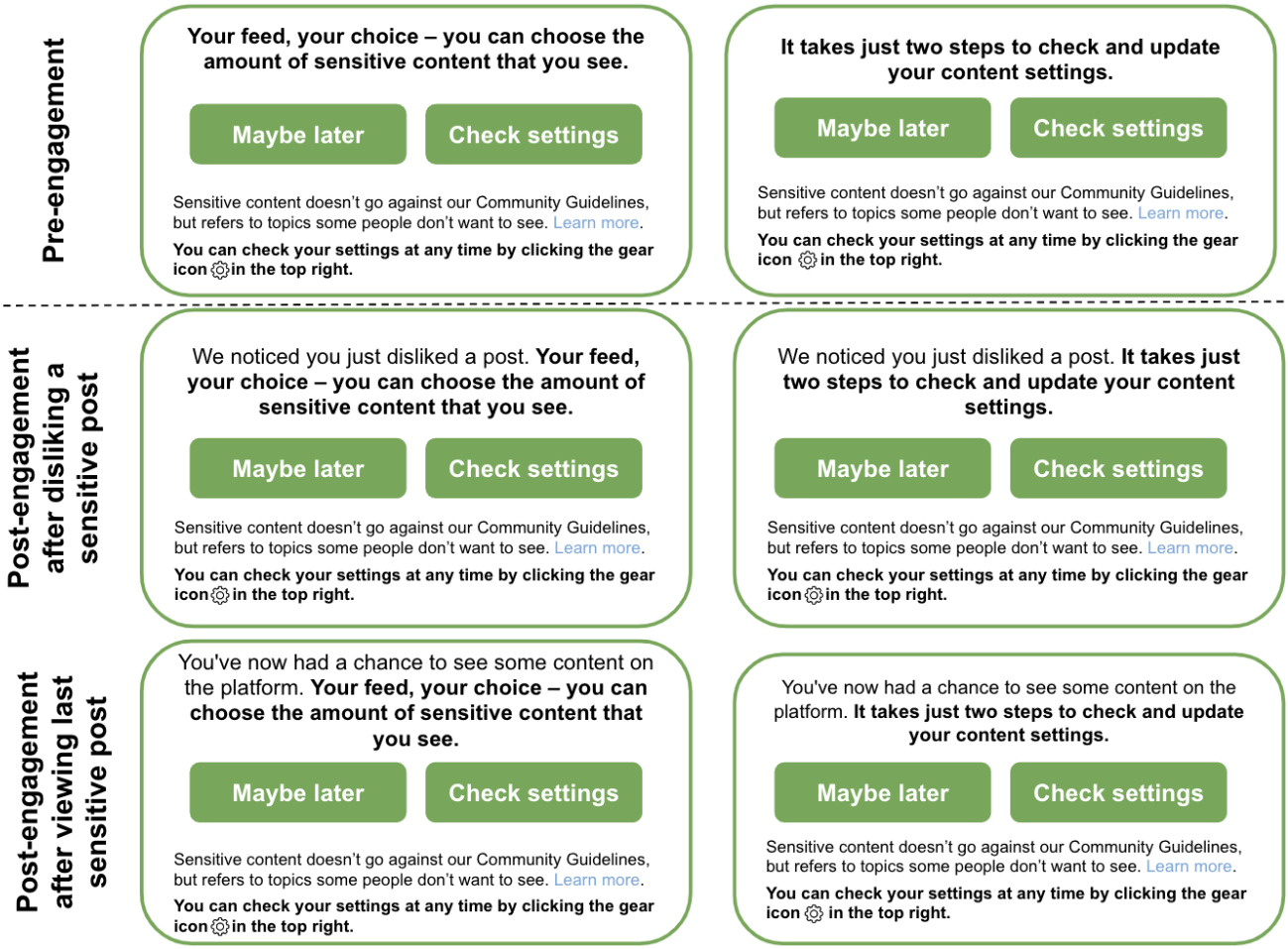
Process message arms

The Process message focused on mitigating concerns that changing content controls is onerous or time consuming.[12]

H5: The probability of reviewing settings at least once in the Process message arms will be significantly higher compared to the Control.

[10] Madden, M., Fox, S., Smith, A., Vitak, J., & Pew Internet & American Life Project, 2007. Digital Footprints: Online identity management and search in the age of transparency. [accessed 28 March 2024].
[11] Centre for Data Ethics and Innovation, 2020. Online targeting: Final report and recommendations. [accessed 25 January 2024].
[12] Centre for Data Ethics and Innovation, 2020. Online targeting: Final report and recommendations. [accessed 25 January 2024].

Figure 5: Empowerment message on the left, Process message on the right

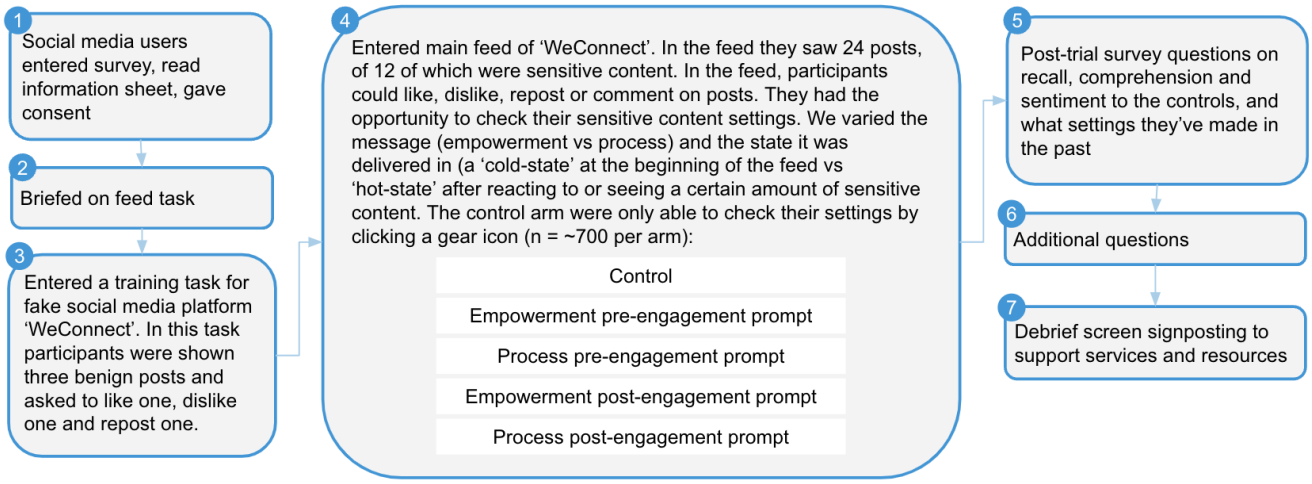


3. Methodology

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 study with an RCT design. In an RCT, research participants are randomly assigned into different groups and exposed to either an intervention or the control. 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 the different interventions have on participants' sentiment, decisions, and behaviours. Figure 6 illustrates the flow of the experiment.

Figure 6: Participant journey.



3.2 Simulated social media platform

3.2.1 Platform design and functionality

We designed our platform, WeConnect, to create a trial environment that mimicked 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 was not based on a single real-world platform, its design was inspired by popular platforms. By making participants' experiences on WeConnect as realistic as possible, we aimed to generate findings that indicated how our interventions would impact users' behaviours on actual platforms.

In the previous trial by Ofcom and BIT using a similar WeConnect platform, 61% of participants said WeConnect was similar or very similar to platforms they had used before and 90% said that the platform was easy or very easy to use.

On WeConnect, participants could scroll through a social media feed, interact with the posts and click on the gear icon to review their content settings. Interaction with the posts on the feed included liking, disliking, reposting, or commenting.

3.2.2 Training task

Before interacting with the main feed, all participants took part in a training task to familiarise themselves with the like, dislike and repost functions. This consisted of a task which involved presenting a feed with 3 non-sensitive posts. Participants had to like, dislike, and repost at least one post before they could proceed with the experiment. As well as familiarising participants with the platform, this training task primed participants to interact with the content.

3.2.3 Main task

After the training task, participants were provided with comprehensive instructions on how to use WeConnect and reminded of the platform functionalities (see Figure 7).

Figure 7: Main feed instructions

Participants had to scroll to the bottom of the feed before they could progress to the next stage of the experiment. After participants scrolled through the feed and clicked 'Next' at the bottom, they progressed to a follow-up survey.

3.2.4 Stimuli

The content consisted of 6 short videos, 6 long videos 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 run by BIT and aimed to keep participants viewing and/or interacting with the feed for 5 minutes. In the training task, participants saw three pieces of non-sensitive content (one short video, one long video, and one short text post).

In the main feed, participants saw 24 pieces of content, including 12 pieces (50%) of sensitive content. The sensitive content categories included in the trial were hate, violence, and misinformation. The non-sensitive posts were made up of non-sensitive content that resembled the type of content users encounter on real social media platforms (see section 3.4 for more detail on content sourcing). The content was presented on the feed in random order, apart from a few restrictions. For example,

participants could not see more than three pieces of sensitive content in a row, and their feed's first and last posts were always non-sensitive.

If participants changed their content setting to "Reduced sensitive content", then the sensitive content in the remainder of the feed was replaced with non-sensitive posts. Sensitive content they had already seen was removed from their feed.

3.2.5 Post-feed survey

After interacting with the main feed, participants completed a post-feed survey, which included questions on recall, their reasons for checking or not checking their settings, their sentiment to the prompt (intervention arms) or the ability to change their settings (control arm) and their past experiences and preferences with content control settings. In the post-feed survey, participants were also asked to report their risk preference using a question based on previous research[13] and adapted to social media platforms. Participants were also asked to retrospectively report their pre-experiment mood[14] and energy[15] (using slider scales adapted from previous research[16] that we updated so response options could be more easily understood). We asked about risk preference, mood, and energy as we speculated that these factors might have a relationship with the likelihood of checking sensitive content settings. We wanted to explore this to inform future research. These psychological measures were included to use as covariates in the analysis. Participants were also asked to provide additional demographic information that we did not receive from the panel, including social grade, and social media platform use.

3.2.6 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 (verbalise their thought processes) as they interacted with the experiment. Think-aloud protocols are a technique commonly used in product design. Participants voiced their thoughts as they went through the platform and experiment, giving 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 understand the definition of sensitive content?).

Based on the researcher's observations and feedback on the platform and content voiced by participants, BIT iteratively updated the design of the platform, interventions, and survey questions. These updates did not lead to any major changes in the overall trial design. However, we changed the format of the prompts from an in-feed post to a pop-up overlaying the content, based on feedback in the user-testing sessions.

---

[13] Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G., 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences, *Journal of the European Economic Association*, 9(3), 522–550.

[14] Pre-experiment mood was measured using a post-feed question asking participants to rate their mood before the experiment on a slider scale with scores ranging from 0 (very negative) - 100 (very positive).

[15] Pre-experiment energy was measured using a post-feed question asking participants to rate their pre-experiment energy on a slider scale with scores ranging from 0 (very low energy) - 100 (very high energy).

[16] Betella, A., & Verschure, P. F., 2016. The affective slider: A digital self-assessment scale for the measurement of human emotions, *PLoS One,*11(2): e0148037.

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
- use/have used a social media platform
- not taken part in the previous trial ran by Ofcom and BIT on a similar platform

3.3.2 Power calculations

The sample size was based on power calculations for our primary outcome (whether the participant reviewed their content settings; see section 3.5.3). In the absence of published online experiments looking at comparable outcomes, we conducted calculations for baseline proportions ranging from 20%-50% (see Table 2), assuming 80% statistical power and a significance level of α = 0.83% (5% / 6; correcting for 6 comparisons in primary analyses[17]). A sample size of 3,500 participants (700 participants per arm) would allow us to detect a minimum detectable effect size of 9.25pp (percentage point difference) between arms where 50% of participants in the baseline arm reviewed their content settings. We deemed this sufficient for an online experiment and consistent with previous online experiments conducted by Ofcom.[18]

Table 2. Power calculations for a sample of 3500 participants (700 per arm) assuming 80% statistical power and a significance level of α = 0.83%.

Outcome baseline	Minimum detectable effect size (% point difference)
20%	7.91%
35%	9.08%
50%	9.25%

3.3.3 Data collection

All participants were recruited through the panel aggregator Lucid. Each participant received financial compensation, with payments being administered by the panel providers they're registered with. Participants were only invited to take part in the experiment by Lucid if they were aged 18 years or older and lived in the UK. We then used a screening question to ensure only participants who use/have used a social media platform with a feed were able to continue with the experiment.

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

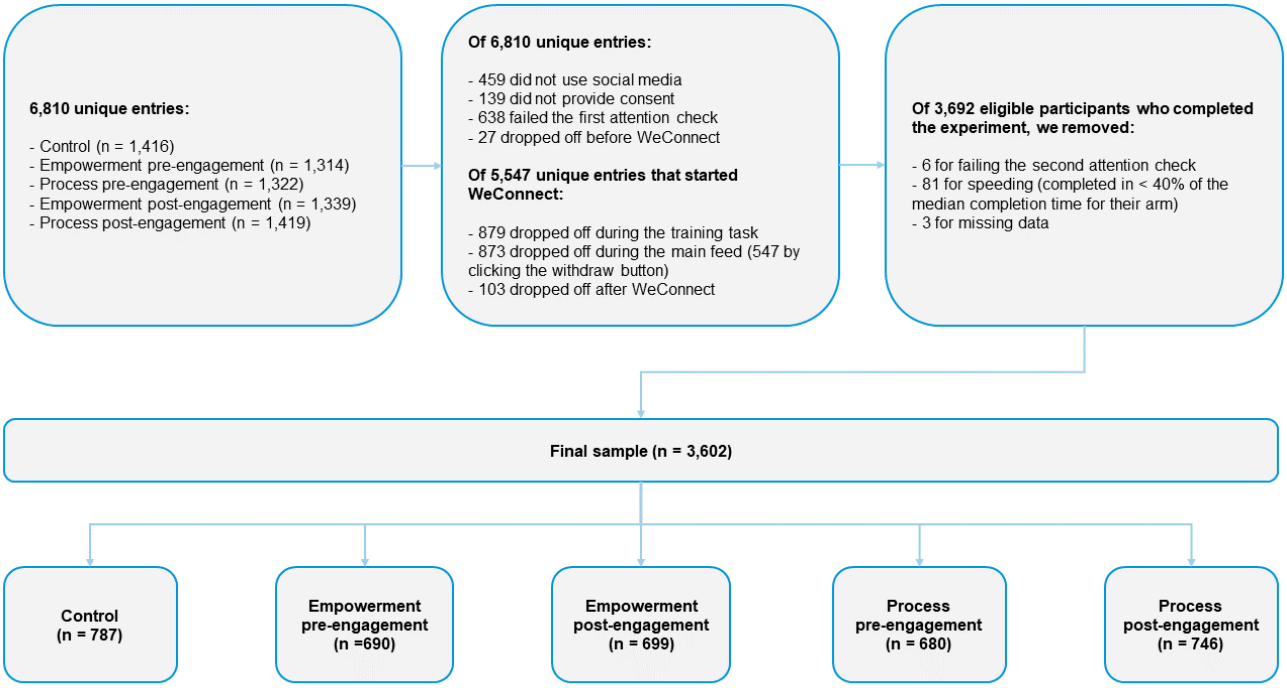
[17] Note that for our analyses we use a Benjamini-Hochberg (BH) correction to adjust for multiple comparisons; however, it is not possible to apply this correction prior to data collection and so for power calculations we use a more conservative Bonferroni correction.
[18] Ofcom, 2023. Boosting users' safety online: Microtutorials.

collected from the participants. Participants were made aware of this 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. There were no data collection issues, so we proceeded to full launch (soft launch data was included in our final sample). During data collection, we continued to monitor the 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 people 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 8 shows the full participant flow with numbers on how many submissions were excluded at which part of the process.

Figure 8: Participant flow diagram.



3.4 Ethical considerations

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

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 and social media platforms

As a result, content types displaying hate, violence and misinformation were included in the trial.

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.[19]

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 was 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 included the 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 time 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.

[19] 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 unique entries to the experiment who passed the social media screener, provided consent, passed the attention check and who made it to or past the WeConnect platform without dropping off (n = 5,547) using a linear regression with the last page of the experiment they completed as the outcome variable and the intervention arm as the predictor variable.

We then checked that our final sample (n = 3,602) was balanced in terms of demographics (age, gender, ethnicity, annual household income (pre-tax), education, urbanicity, employment, region, social grade, and social media platform use) across intervention 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 analyses based on an agreed upon hierarchy. We used a significance level of 5% throughout all analyses, correcting for multiple comparisons separately within the primary analysis (6 comparisons) and across the secondary analyses (6 comparisons) using the Benjamini-Hochberg adjustment.

3.5.3 Primary analysis

The primary outcome was whether the participant checked their content settings and we conducted two multivariate logistic regressions on this outcome.

Primary analysis regression 1

$$check_i \sim bernoulli\,(p_i)\,;\; logit(p_i) \;=\; \alpha \;+\; \beta_1 message_i + \beta_2 covariates_i$$

- *check$_i$ is* a binary variable for whether participant *i* checked their content settings at least once during the experiment (coded as 1) or not at all (coded as 0), across checks via the prompt (treatment arms only) or via the gear icon (all arms). We excluded pre-prompt checks via the gear icon for the intervention arms because these checks would not have been a result of the interventions.
- *message$_i$* is a dummy coded variable for whether participant *i* was exposed to an Empowerment prompt or Process prompt, with the Control arm as the reference level.
- *covariates$_i$* is a vector of covariates. Categorical covariates included age (18-24; 25-54; 55 and over), gender (male; female; other), ethnicity (White; Asian; Black; Mixed or other), annual pre-tax income (£40,000 or over; less than £40,000), and education (degree; no

degree; prefer not to say). Continuous covariates included platform use,[20] risk preference[21] (measured with a post-feed question based on previous research[22] and adapted to social media platforms based on feedback from user testing), and retrospectively reported pre-experiment mood[23] and pre-experiment energy[24] (using slider scales adapted from previous research[25] that we updated so response options could be more easily understood).

- Primary analysis regression 1 was conducted on the full sample (n = 3,602) and we made the following three comparisons:
    - Control arm vs. Empowerment arms (across Pre- and Post-engagement arms)
    - Control arm vs. Process arms (across Pre- and Post-engagement arms)
    - Empowerment arms vs. Process arms (across Pre- and Post-engagement arms)

Primary analysis regression 2

$$check_i \sim bernoulli\,(p_i)\,;\; logit(p_i) \;=\; \alpha \;+\; \beta_1 timing_i \;+\; \beta_2 covariates_i$$

- $check_i$ as specified in primary analysis regression 1.
- pre_i is a dummy coded variable for whether participant *i* was exposed to a Pre-engagement prompt or a Post-engagement prompt, with the Control arm as the reference level.
- $covariates_i$ as specified in primary analysis regression 1.
- Primary analysis regression 2 was conducted on the full sample (n = 3602) and we made the following three comparisons:
    - Control arm vs. Pre-engagement arms (across Empowerment and Process arms)
    - Control arm vs. Post-engagement arms (across Empowerment and Process arms)
    - Pre-engagement arms vs. Post-engagement arms (across Empowerment and Process arms)

3.5.4 Secondary analyses

Secondary analysis 1

For secondary analysis 1, we used the same outcome as the primary outcome, but excluded the Control arm and compared individual intervention arms against each other rather than pooling them together depending on the message or timing of the prompt.

---

[20] Platform use was measured in a post-feed survey question asking participants if they had accounts with any of the following platforms: TikTok, Instagram, Facebook, YouTube, Snapchat, Twitter/X, Reddit, LinkedIn, OnlyFans, Vimeo, and WhatsApp. For each platform, if the participant did not have an account then they were coded as 0. If they did have an account with the platform, then they were asked how often they use the platform ("I don't use this anymore" coded as 0; "Less often" coded as 1; "Once a week" coded as 2; "Several times a week" coded as 3; "Once a day" coded as 4; "Several times a day" coded as 5). For each participant, we calculated a platform usage score by summing the scores across platforms.

[21] Risk preference was measured in a post-feed survey asking participants to think about the content they see on social media platforms and rate their risk preference on a scale from 0 ("Not at all willing to take risks") to 10 ("Very willing to take risks").

[22] Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G., 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences, *Journal of the European Economic Association*, 9(3), 522–550.

[23] Pre-experiment mood was measured using a post-feed question asking participants to rate their mood before the experiment on a slider scale with scores ranging from 0 (very negative) - 100 (very positive).

[24] Pre-experiment energy was measured using a post-feed question asking participants to rate their pre-experiment energy on a slider scale with scores ranging from 0 (very low energy) - 100 (very high energy).

[25] Betella, A., & Verschure, P. F., 2016. The affective slider: A digital self-assessment scale for the measurement of human emotions, *PLoS One*,11(2): e0148037.

Secondary analysis 1 regression 1

$$check_i \sim bernoulli\ (p_i)\ ;\ logit(p_i)\ =\ \alpha\ +\ \beta_1 treatment_i\ +\ \beta_2 covariates_i$$

- *check*$_i$ as specified in primary analysis regression 1.
- *treatment*$_i$ is a dummy coded variable for whether participant *i* was exposed to any of the treatment arms (Empowerment Pre-engagement prompt, the Process Pre-engagement prompt, the Empowerment Post-engagement prompts or the Process Post-engagement prompt).
- *covariates*$_i$ as specified in primary analysis regression 1.
- Secondary analysis 1 regression 1 was conducted on the full sample excluding the Control arm (n = 2,815) and we made the following four comparisons:
    - Empowerment Pre-engagement vs. Empowerment Post-engagement arms
    - Process Pre-engagement vs. Process Post-engagement arms
    - Empowerment Pre-engagement vs. Process Pre-engagement arms
    - Empowerment Post-engagement vs. Process Post-engagement arms

Secondary analysis 2

The secondary analysis 2 outcome was whether the participant's final content settings choice after interacting with the feed matched their self-reported usual content preferences collected in the post-feed survey.

Secondary analysis 2 regression 1

$$match_i \sim bernoulli\ (p_i)\ ;\ logit(p_i)\ =\ \alpha\ +\ \beta_1 process_i\ +\ \beta_2 covariates_i$$

- *match*$_i$ is a binary variable for whether the final content settings choice of participant *i* matched (coded as 1) or mismatched (coded as 0) their self-reported usual content preferences collected in the post-feed survey.[26]
- *process*$_i$ is a binary variable for whether participant *i* was exposed to a Process prompt (Process Pre-engagement or Process Post-engagement arms; coded as 1) or not (coded as 0).
- *covariates*$_i$ as specified in primary analysis regression 1.
- Secondary analysis 2 regression 1 was conducted on the full sample excluding the Control arm (n = 2,815) and we made the following comparison:
    - Empowerment vs. Process arms (across Pre- and Post-engagement arms)

Secondary analysis 2 regression 2

$$match_i \sim bernoulli\ (p_i)\ ;\ logit(p_i)\ =\ \alpha\ +\ \beta_1 post_i\ +\ \beta_2 covariates_i$$

- *match*$_i$ as specified in secondary analysis 2 regression 1.
- post$_i$ is a binary variable for whether participant *i* was exposed to a Post-engagement prompt (Empowerment Post-engagement or Process Post-engagement; coded as 1) or not (coded as 0).

[26] We recorded the participant's final content settings choice after the participant finished interacting with the feed ("All content" vs. "Reduced sensitive content"). In a post-feed survey, we asked the participant how much sensitive content they are usually comfortable seeing on their social media feed ("All content" vs. "Reduced sensitive content" vs. "Don't know"). For secondary outcome 2, we coded whether the participant's final content settings choice after interacting with the feed matched (coded as 1) or mismatched (coded as 0) their self-reported usual content preferences collected in the post-feed survey. If a participant responded with "Don't know" in the post-feed survey then they were coded as mismatches, because a "Don't know" response would not indicate a confident match.

- $covariates_i$ as specified in primary analysis regression 1.
- Secondary analysis 2 regression 2 was conducted on the full sample excluding the Control arm (n = 2,815) and we made the following comparison:
    - Pre-engagement vs. Post-engagement arms (across Empowerment and Process arms)

3.5.5 Exploratory analyses

For the following exploratory outcomes, unless otherwise stated, we used the same model specifications as secondary analysis 2, but replaced the outcome variable and did not correct for multiple comparisons.

Exploratory analysis 1

- $recall_i$ is a binary outcome variable for whether participant *i* correctly identified in a post-feed survey that they could change their content settings[27] (coded as 1) or not (coded as 0).

Exploratory analysis 2

- $final_i$ is a binary outcome variable for whether the final content settings choice made by participant *i* after interacting with the feed was "Reduced sensitive content" (coded as 1 or "All content types" (coded as 0).

Exploratory analysis 3

- $reviewCount_i$ is a count outcome variable for the number of times participant *i* reviewed their settings, across checks via the prompt (treatment arms only) or via the gear icon (all arms). We excluded pre-prompt checks via the gear icon for the treatment arms because these checks would not have been a result of the interventions.
- Instead of a logistic regression we conducted zero-inflated regressions (after checking for excess zeros and overdispersion).

Exploratory analysis 4

- $messageTime_i$ is a continuous outcome variable for the length of time in seconds that participant *i* spent viewing the intervention prompt.
- Instead of a logistic regression we conducted a linear regression.

Exploratory analysis 5

- $reviewTime_i$ is a continuous outcome variable for the length of time in seconds that participant *i* spent viewing the review prompt (asking participants to confirm their content choice).
- Instead of a logistic regression we conducted a linear regression.

[27] Recall was measured using a single choice question where participants were asked which settings, they could review in the WeConnect feed (content settings, privacy settings, location-sharing settings, or language settings, including a don't know option).

Exploratory analysis 6

- *sentimentEasy*$_i$ is a binary outcome variable measured in a post-feed survey indicating the extent to which participant *i* felt the prompt was easy to understand. Participants could answer "Not at all" (coded as 0), "A little" (coded as 0), "Moderately" (coded as 1), and "Very much" (coded as 1).

Exploratory analysis 7

- *sentimentControl*$_i$ is a binary outcome variable measured in a post-feed survey indicating the extent to which participant *i* felt the prompt made them feel in control of the content they saw on WeConnect. Participants could answer "Not at all" (coded as 0), "A little" (coded as 0), "Moderately" (coded as 1), and "Very much" (coded as 1).

Exploratory analysis 8

- *sentimentAnnoy*$_i$ is a binary outcome variable measured in a post-feed survey indicating the extent to which participant *i* felt the prompt was annoying. Participants could answer "Not at all" (coded as 0), "A little" (coded as 0), "Moderately" (coded as 1), and "Very much" (coded as 1).

Exploratory analysis 9

- *sentimentExpect*$_i$ is a binary outcome variable measured in a post-feed survey indicating the extent to which participant *i* felt the prompt was something they would expect to see when scrolling through a social media website. Participants could answer "Not at all" (coded as 0), "A little" (coded as 0), "Moderately" (coded as 1), and "Very much" (coded as 1).

Exploratory analysis 10

- *sentimentUseful*$_i$ is a binary outcome variable measured in a post-feed survey indicating the extent to which participant *i* felt the prompt was a useful reminder. Participants could answer "Not at all" (coded as 0), "A little" (coded as 0), "Moderately" (coded as 1), and "Very much" (coded as 1).

For the following exploratory analyses, we tested the association between the primary outcome (*check*$_i$) and the psychological variables (risk preference, and retrospectively reported pre-experiment mood and pre-experiment energy) across treatment arms. In these analyses, the predictor was continuous and so instead of multiple comparisons we tested whether the predictor was associated with our outcome.

Exploratory analysis 11

$$check_i \sim bernoulli\,(p_i)\,;\; logit(p_i) \;=\; \alpha \;+\; \beta_1 riskPreference_i \;+\; \beta_2 covariates_i$$

- *check*$_i$ as specified in primary analysis regression 1.
- *riskPreference*$_i$ as specified in *covariates*$_i$ for primary analysis regression 1.
- *covariates*$_i$ as specified in primary analysis regression 1 but not including risk preference.

- Exploratory analysis 11 was conducted on the full sample (n = 3,602) and we tested the association between risk preference and whether the participant checked their content settings.

Exploratory analysis 12

$$check_i \sim bernoulli\ (p_i)\ ;\ logit(p_i)\ =\ \alpha\ +\ \beta_1 mood_i\ +\ \beta_2 covariates_i$$

- *check*$_i$ as specified in primary analysis regression 1.
- *mood*$_i$ as specified in *covariates*$_i$ for primary analysis regression 1.
- *covariates*$_i$ as specified in primary analysis regression 1 but not including mood.
- Exploratory analysis 12 was conducted on the full sample (n = 3,602) and we tested the association between pre-experiment mood and whether the participant checked their content settings.

Exploratory analysis 13

$$check_i \sim bernoulli\ (p_i)\ ;\ logit(p_i)\ =\ \alpha\ +\ \beta_1 energy_i\ +\ \beta_2 covariates_i$$

- *check*$_i$ as specified in primary analysis regression 1.
- *energy*$_i$ as specified in *covariates*$_i$ for primary analysis regression 1.
- *covariates*$_i$ as specified in primary analysis regression 1 but not including energy.
- Exploratory analysis 13 was conducted on the full sample (n = 3,602) and we tested the association between pre-experiment energy and whether the participant checked their content settings.

3.5.6 Sensitivity analyses

We conducted three sensitivity analyses. First, we reran the primary analysis and secondary analysis 1, but without excluding pre-prompt checks via the gear icon for the treatment arms. This sensitivity check was not pre-specified and was conducted to ensure that results were consistent when the outcome variable *check*$_i$ was coded in the same way for control and treatment arms i.e. including all checks via the gear icon. Second, we conducted a pre-specified sensitivity check and reran the primary and secondary analyses, but excluding participants from the Post-engagement arms who did not dislike any sensitive posts. Third, we conducted a pre-specified sensitivity check and reran exploratory analyses 6-10 but using ordinal regressions with the outcomes coded as 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).

4. Results

4.1 Sample characteristics

We found evidence for an overall effect of differential attrition (adjusted R^2 = 0.002811, F(4, 4663) = 3.286, p = .011). The experiment consisted of 18 separate screens that participants had to progress through in order, and participants who made it to screen 18 were considered to have completed the experiment. Of participants who made it to the main task of the WeConnect feed (screen 6 of the 18 experiment screens; n = 5,547), participants in the Control arm had a mean last experiment screen of 16.09 and participants in each of the treatment arms were consistently significantly more likely to drop out of the experiment compared to the Control arm (Empowerment Pre-engagement, β = -0.65, p < .01; Process Pre-engagement, β = -0.55, p < .05; Empowerment Post-engagement, β = -0.69, p < .01; Process Post-engagement, β = -0.52, p < .05; *N* who completed the experiment: Control = 787, Empowerment Pre-engagement = 690, Process Pre-engagement = 680, Empowerment Post-engagement = 699, Process Post-engagement = 746). One possible explanation for this result is that the prompts were perceived by some participants as technical errors with the survey which increased the likelihood of dropping off.

The sample was balanced across treatment arms for all variables (all p > .05), except for platform use, (F(4) = 2.539, p = .0381; Mean platform use score: Control = 19.11, Empowerment Pre-engagement = 18.53, Process Pre-engagement = 18.87, Empowerment Post-engagement = 19.55, Process Post-engagement = 19.84). Despite this, by including platform use as covariate in all statistical models as planned, the effects of this imbalance was minimal. Since the sample was generally balanced on demographics and the differential attrition was consistent across treatment arms, we continued with our prespecified analysis plan.

The demographics for our final sample (n = 3,602) are reported in Table 3.

Table 3. Sample demographics for final sample (n = 3,602)

	Category	% of the sample
Age	18-24	12%
	25-54	62%
	55 and over	26%
Gender	Male	46%
	Female	54%
	Other (e.g. non binary)	< 1%
Ethnicity	White	87%
	Asian	6%
	Black	4%
	Mixed or other	3%
Annual pre-tax income	£40,000 or over	47%
	Less than £40,000	53%
Education	Degree	32%
	No degree	65%
	Prefer not to say	3%
Urbanicity	Urban	29%
	Suburban	48%
	Rural	23%
Employed	Employed	74%
	Unemployed	3%
	Inactive	23%
Location	London	12%
	Midlands	17%
	North	26%
	South & East	32%
	Wales, Scotland & Northern Ireland	14%
Social grade	High	35%
	Medium	57%
	Low	7%
	Don't know	< 1%
Psychological metrics	Energy just before the experiment (0 Very low energy - 100 Very high energy)	Mean = 60.25 SD = 22.43
	Mood just before the experiment (0 Very negative - 100 Very positive)	Mean = 67.01 SD = 20.27
	Willingness to take risk (0 Not at all willing - 10 Very willing)	Mean = 5.36 SD = 2.68

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

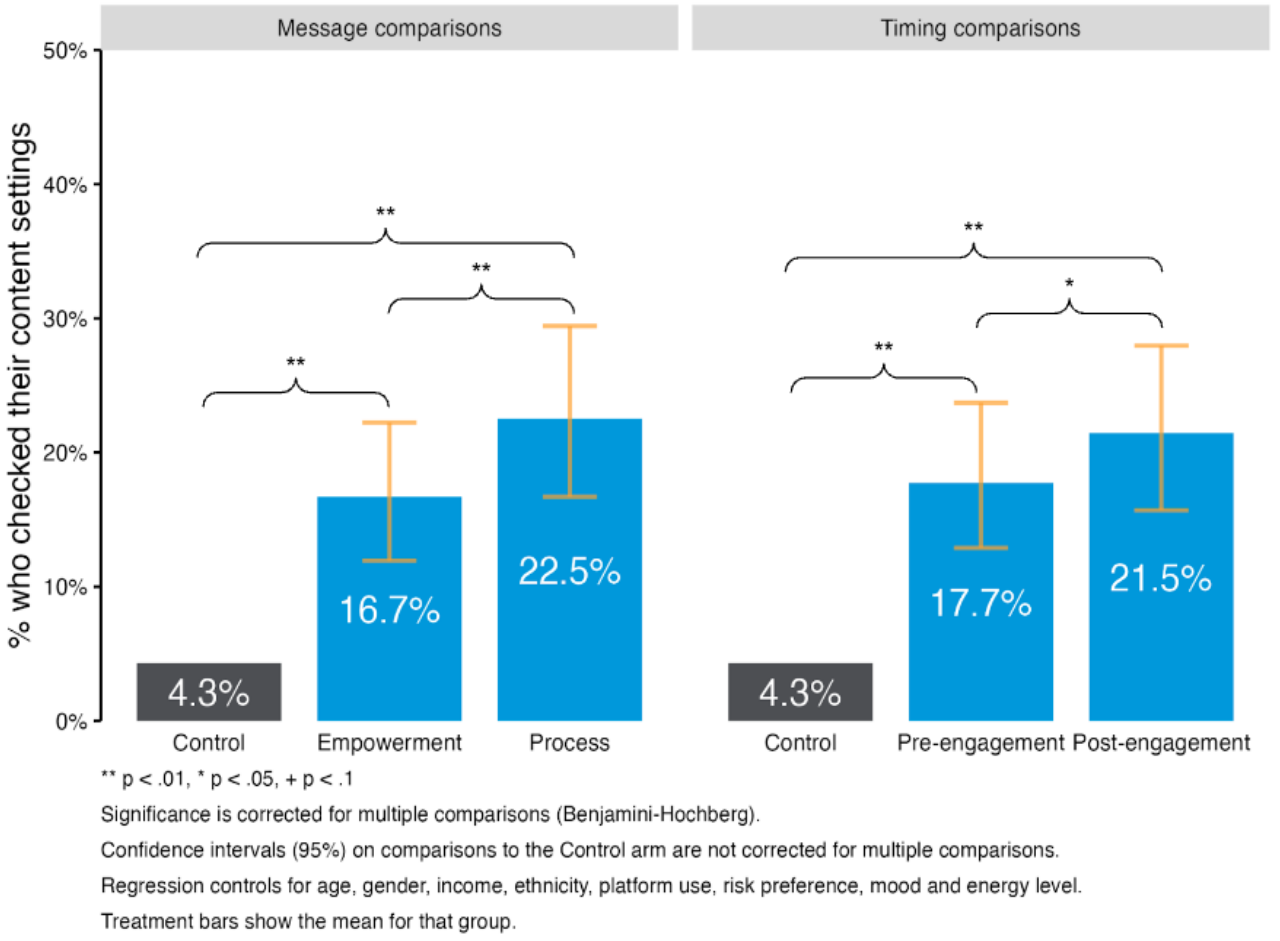
4.2 Primary analysis: Whether people checked their content settings

Participants who saw the Empowerment message were significantly more likely to check their content settings than those in the Control arm, who saw no prompt at all $p < .01$ (16.7% compared to 4.3%). Participants who saw the Process message were also significantly more likely to check their content settings than those in the Control arm, $p < .01$ (22.5% compared to 4.3%). Additionally, participants who saw the Process messages were significantly more likely to check their settings than those who saw the Empowerment messages, $p < .01$.

Participants in the Pre-engagement and Post-engagement arms were significantly more likely to check their settings than those in the Control arm, who didn't see any prompt, $p < .01$ (17.7% and 21.5% respectively, compared to 4.3% in the Control arm). Participants in the Post-engagement arms

were significantly more likely to check their settings than those in the Pre-engagement arms, $p < .05$. Results are shown in Figure 9.

Figure 9: The results of the primary analysis on the percentage of participants who checked their content setting.



As a sensitivity check, we conducted the same analysis coding participants who only checked their settings before seeing a prompt (n = 15) as 1 to analyse the control arm in the same way as the treatment arms and therefore exclude the possibility that pre-intervention behaviour introduced unobserved factors that influenced the outcome. We found the same results. Participants who saw the Empowerment message and participants who saw the Process message were significantly more likely to check their settings than those in the Control arm who saw no prompt, both $p < .01$ (17.4% for the Empowerment message and 22.9% for the Process message vs 4.3%). Participants who saw the Process message were significantly more likely to check their settings than those who saw the Empowerment message, $p < .01$. Participants in the Pre-engagement arms and Post-engagement arms were significantly more likely to check their settings than those in the Control arm who didn't see any prompt, both $p < .01$ (18.0% for Pre-engagement arms and 22.3% for Post-engagement arms vs 4.3%). Participants in the Post-engagement arms were significantly more likely to check their settings than those in the Pre-engagement arms, $p < .01$.

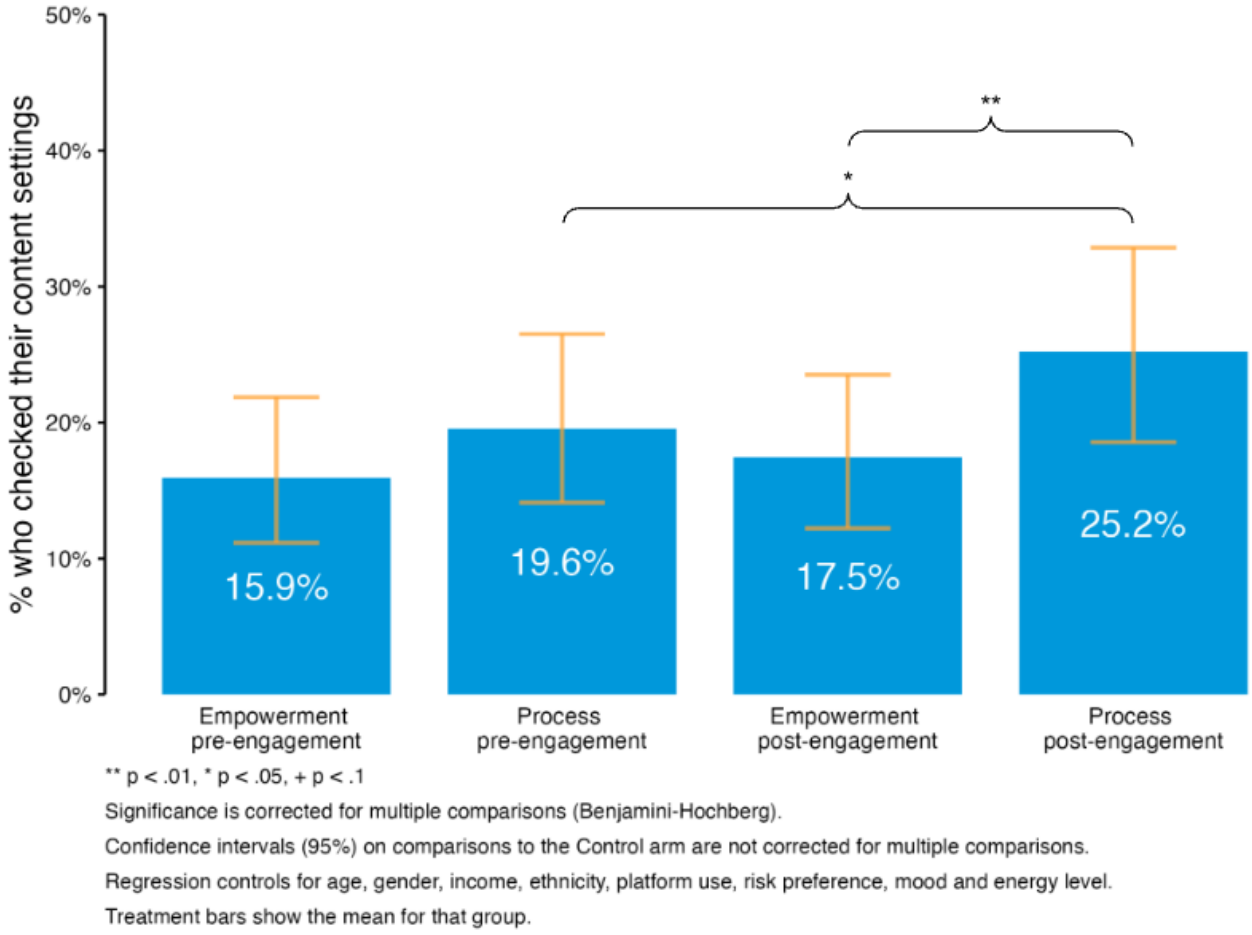
4.3 Secondary analyses

4.3.1 Further comparisons on the primary outcome

Because the treatment arms performed significantly better than the Control in the primary analysis, we will compare the individual treatment arms to each other and not to the Control.

Participants who saw the Process message after engaging with content were significantly more likely to check their settings than those who saw the same message before engaging with content, $p < .05$ (25.2% vs. 19.6% respectively). Participants who saw the Process message after engaging with content were significantly more likely to check their content settings than those who saw the Empowerment message at the same time, $p < .01$ (25.2% vs. 17.5% respectively). There were no significant differences between the Empowerment Pre-engagement and Process Pre-engagement or Empowerment Post-engagement arms. Results are shown in Figure 10.

Figure 10: The results of secondary analysis 1 on the percentage of participants who checked their content settings in each treatment arm.



4.3.2 Whether their settings match their preferences

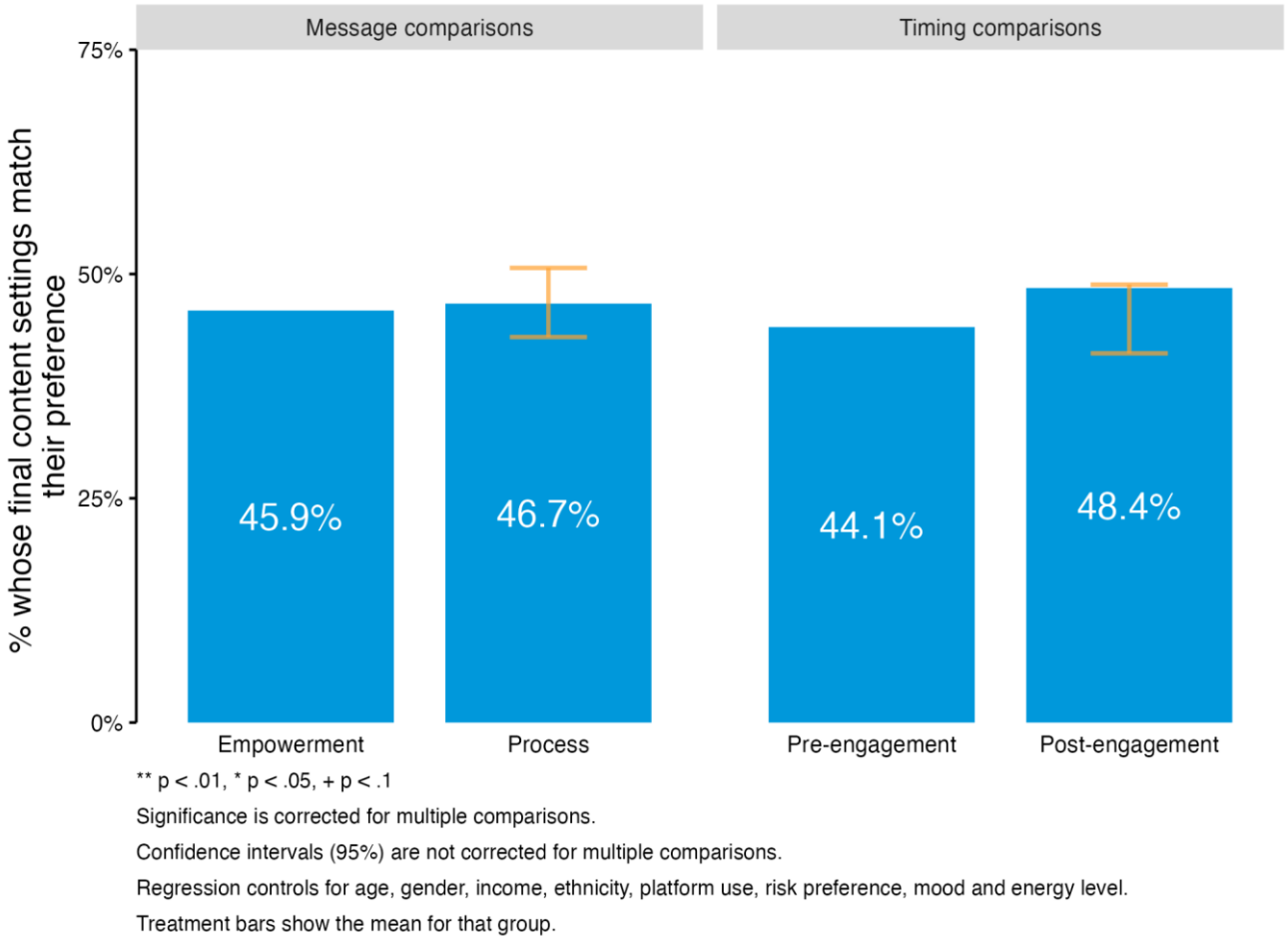
Overall, 8.3% of participants ended the feed task with their content settings set to show reduced sensitive content (see section 4.4.2 for details). At the same time, 49% said that they're usually

comfortable seeing reduced sensitive content on their feed and 39% said they're usually comfortable seeing all content types (12% said they don't know).

As a result, the final content settings at the end of the feed matched their usual preference for 44.1% of participants and did not match for another 44%. This was not significantly different for participants who saw the Empowerment message compared to the Process message.

There were also no significant differences between participants who saw the prompt before engaging with content compared to those who saw it after engaging with content. Results are shown in Figure 11.

Figure 11: The results of secondary analysis 2 on the percentage of participants whose final content settings match their usual preference.



For post-hoc analysis, we compared the treatment arms against the Control arm, where 36.1% of participants' final content settings matched their preferences (see Table 4). Participants in both the Empowerment message and Process message were significantly more likely to have content settings that matched their preference, both $p < .01$. The same was true for participants who saw the prompt before or after engaging with content, both $p < .01$. This suggests that any prompt increased the likelihood of participants having their preferred content settings, but the message or timing did not significantly impact this alignment.

Table 4. The results of post hoc analysis on the percentage of participants whose final content settings match their usual preference, comparing treatment arms against the Control arm.

Outcome	Control	Empowerment	Process	Pre-engagement	Post-engagement
% whose final content settings match their preference	36.1%	45.9% [42.4%-51.8%]**	46.7% [43.3%-52.7%]**	44.1% [40.8%-50.2%]**	48.4% [44.8%-54.1%]**

** p < .01, * p < .05, + < .1
This table reports the means and results of two regressions; one regression comparing Empowerment and Process messages against the control arm and one regression comparing Pre-engagement and Post-engagement timings against the control arm.
Regressions control for age, gender, income, ethnicity, platform use, risk preference, mood, and energy level.
Significance and confidence intervals (95%; reported in brackets) are not corrected for multiple comparisons.

4.4 Exploratory analyses

Note that exploratory analyses have not been corrected for multiple 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'). The decision not to 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. Therefore, the findings in this section should be taken as exploratory rather than hypothesis confirming.

4.4.1 Recall

Across all arms, 35.7% of the full sample correctly recalled that they could change their content settings on WeConnect. There were no significant differences in recall between participants who saw the Empowerment message compared to those who saw the Process message ($p > .05$). Participants who saw the prompt after engaging with content were significantly more likely to correctly recall that they can change their content settings than those who saw the prompt before engaging with content, $p < .01$ (45.5% vs. 33.5% respectively). Results are shown in Table 5.

Table 5. The results of exploratory analyses 1 and 3-10, comparing Empowerment vs. Process messages and Pre-engagement vs. Post-engagement timings.

	Message comparisons		Timing comparisons	
Outcome	**Empowerment**	**Process**	**Pre-engagement**	**Post-engagement**
% who recalled they could change their content settings	40.0%	39.3% [35.6%-42.9%]	33.5%	45.5% [41.5%-49.2%]**
Mean number of times participants checked their settings	0.19	0.27 [0.19-0.62]*	0.21	0.25 [-0.29-0.28]
Mean time spent viewing the intervention prompt (s)	7.38	7.87 [7.34-8.53]+	6.97	8.25 [7.71-8.90]**
Mean time spent viewing the review prompt (s)	0.64	1.02 [0.83-1.22]**	0.74	0.93 [0.72-1.11]+
% who said the prompt was easy to understand	76.2%	74.5% [71.0%-77.7%]	74.0%	76.6% [72.3%-78.8%]
% who said the prompt made them feel in control of the content they saw	65.4%	64.2% [60.9%-68.2%]	62.6%	66.9% [62.5%-69.6%]+
% who thought the prompt was annoying	18.1%	21.9% [18.6%-25.0%]*	21.4%	18.8% [16.2%-21.9%]
% who said the prompt was something they'd expect to see on social media	50.1%	46.1% [42.3%-49.8%]*	49.1%	47.1% [42.7%-50.2%]
% who said the prompt was a useful reminder	63.4%	60.3% [57.0%-64.4%]	60.2%	63.3% [58.9%-66.1%]

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

This table reports the results of two regressions for each outcome; one regression comparing Empowerment vs. Process messages and one regression comparing Pre-engagement vs. Post-engagement timings.

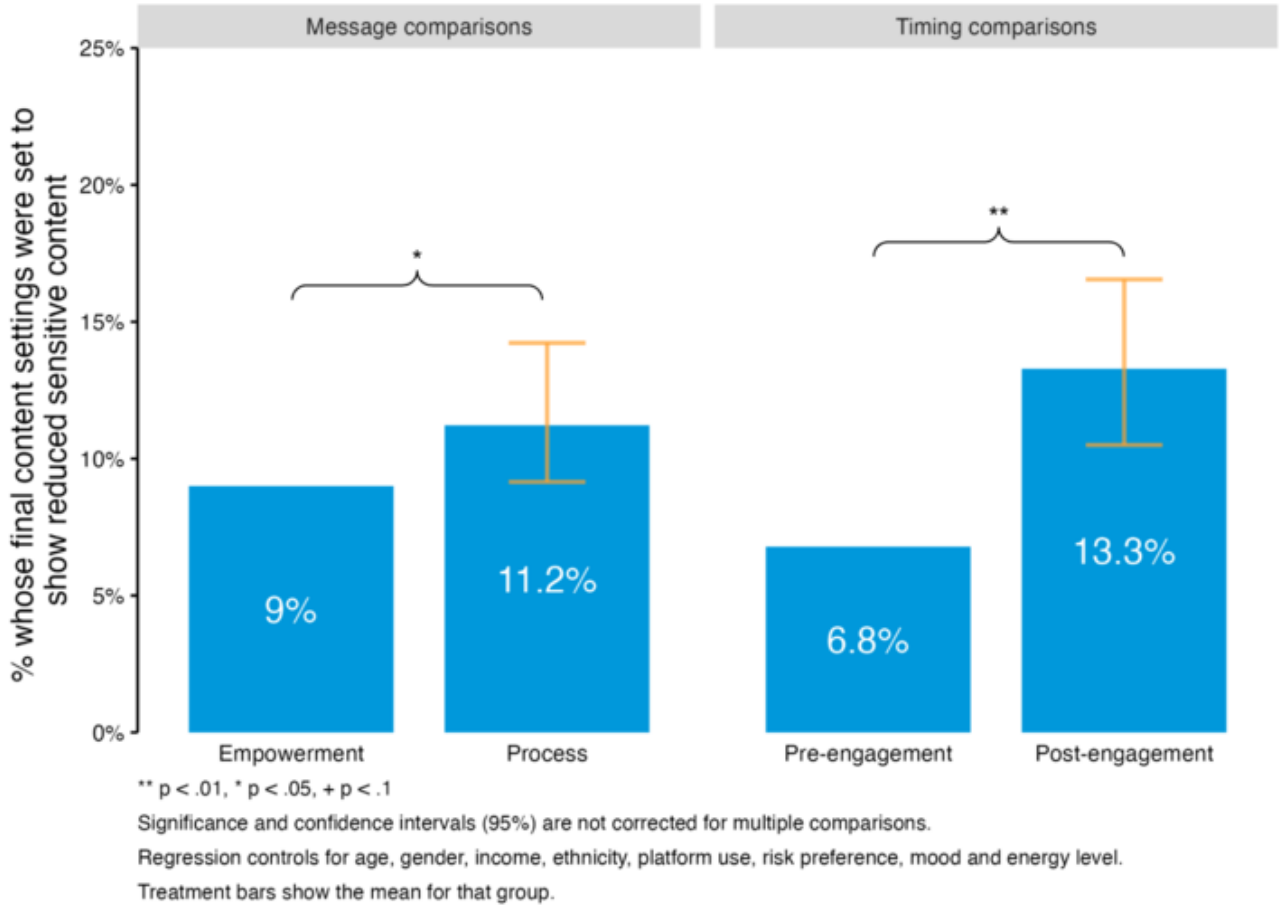
Regressions exclude the control arm and control for age, gender, income, ethnicity, platform use, risk preference, mood, and energy level.
Significance and confidence intervals (95%; reported in brackets) are not corrected for multiple comparisons.

4.4.2 Final choice

Overall, 8.3% of participants ended the feed task with their content settings set to show reduced sensitive content. Participants who saw the Process message were significantly more likely to set their final choice to reduced sensitive content than those who saw the Empowerment message, $p < .05$ (11.2% vs. 9% respectively).

Participants who saw the prompt after engaging with content were significantly more likely to end the feed task seeing reduced sensitive content than those who saw the prompt before engaging with the content, $p < .01$ (13.3% vs. 6.8% respectively). Results are shown in Figure 12.

Figure 12: The results of exploratory analysis 2 on the percentage of participants who ended the feed with their content setting set to show reduced sensitive content.



4.4.3 How many times participants reviewed their settings

3,015 participants did not review their settings during the experiment. 503 participants reviewed their settings once, 78 participants reviewed their settings twice, and 6 participants reviewed their settings three times. No one checked their settings more than three times. Because the values were

overdispersed (X^2 (3586) = 2967.73, dispersion = 1.060 on message comparison; X^2 (3586) = 2985.42, dispersion = 1.067 on timing comparison) a zero-inflated Poisson regression was used to analyse the results of how many times participants reviewed their settings.

Participants who saw the Process message reviewed their settings significantly more often than those who saw the Empowerment message, $p < .05$ (mean of 0.19 times vs. 0.27 times respectively). Results are shown in Table 5.

There were no significant differences in the number of times participants checked their setting between participants who saw the prompt before engaging with content and after engaging with content. Results are shown in Table 5.

4.4.4 Time viewing prompt

The prompt was open for an average of 7.6 seconds. There were no significant differences in the amount of time the Empowerment message was open for compared to the Process message. Results are shown in Table 5.

Participants who saw the prompt after engaging with content had it open for significantly longer than those who saw it before engaging with content, $p < .01$ (8.25 seconds vs. 6.97 seconds respectively). Results are shown in Table 5.

4.4.5 Time reviewing

Participants who saw the Process message spent longer reviewing their settings than those who saw the Empowerment message, $p < .05$ (0.64 seconds vs. 1.02 second respectively).

There were no significant differences in time spent reviewing their settings between those who saw the prompt after engaging with content and those who saw it before engaging with content. Results are shown in Table 5.

4.4.6 Sentiment

Ease of understanding

75.3% said the prompt was easy to understand. There were no significant differences between participants who saw the Empowerment or process prompt. There were also no significant differences in ease of understanding for participants who saw the prompt before engaging with content and those who saw it after.

Feeling of control

64.8% said the prompt made them feel in control of the content they'll see on WeConnect. There were no significant differences between participants who saw the Empowerment or Process prompt.

There were also no significant differences in feeling of control for those who saw the prompt before engaging with content and those who saw it after. In our sensitivity analysis, when analysing the same data but using an ordinal logistic regression instead of a binary logistic regression, there was a significant difference between those who saw the prompt before engaging with content and those who saw it after engaging with content, $p < .01$. A Brant test showed that the proportional odds assumption holds (X^2 (30) = 38.79, $p = .13$).

Annoying

Participants who saw the Process message were significantly more likely to think it was annoying than those who saw the Empowerment message, $p < .05$ (21.9% vs. 18% respectively).

There were no significant differences in whether the participant found the prompt annoying for those who saw the prompt before engaging with content and those who saw it after. In our sensitivity analysis, when analysing the same data but using an ordinal logistic regression instead of a binary logistic regression, there was a significant difference between those who saw the prompt before engaging with content and those who saw it after engaging with content, $p < .01$. However, the Brant test showed that the proportional odds assumption was violated (X^2 (30) = 48.74, $p = .02$), so these results should be interpreted with caution.

Something they'd expect to see

Significantly more participants who saw the Empowerment message said the prompt is something they'd expect to see on social media than those who saw the Process message, $p < .05$ (50.1% vs 46.1% respectively). It is 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 arms. However, a Brant test showed that the proportional odds assumption was violated (X^2 (30) = 45.91, $p = .03$), so these results should be interpreted with caution.

There were no significant differences in whether the prompt was something they'd expect to see on social media between participants who saw the prompt before engaging with content or after.

Useful

Overall, 61.8% of participants said the prompt was a useful reminder. There were no significant differences in this between participants who saw the Empowerment prompt compared to those who saw the process prompt.

There were also no significant differences in whether the participant thought the prompt was a useful reminder for those who saw the prompt before engaging with content and those who saw it after. In our sensitivity analysis, when analysing the same data but using an ordinal logistic regression instead of a binary logistic regression, there was a significant difference between those who saw the prompt before engaging with content and those who saw it after engaging with content, $p < .01$. However, a Brant test showed that the proportional odds assumption was violated (X^2 (30) = 54.43, $p < .00$), so these results should be interpreted with caution.

Detailed results

Results from the logistic regressions on binarised sentiment outcomes are presented in Table 5. Results from the sensitivity analyses using ordinal regressions on ordinal sentiment outcomes are presented in Table 6. Results from the Brant test of the proportional odds assumption are presented in the Annex.

The discrepancy between the binary and the ordinal sensitivity analyses may be because of the small numbers of participants who selected "Not at all" or "Very much" for certain questions (see Table 6). This would not have affected the binary logistic analysis as in that analysis the categories "Not at all" and "A little" were collapsed and the categories "Moderately" and "Very much" were collapsed.

However, in the ordinal logistic analysis this category was considered in the model separately. Having very few observations in that category increases the uncertainty in our results, as there is less data to go by, which can negatively affect statistical power. 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.

Table 6. Results of ordinal regression on secondary outcomes 6-10.

	Empowerment				Process				
Outcome	**Not at all**	**A little**	**Moderately**	**Very much**	**Not at all**	**A little**	**Moderately**	**Very much**	***p***
% who said the prompt was easy to understand	4.4%	19.4%	29.5%	46.7%	5.7%	19.9%	28.4%	46.1%	
% who said the prompt made them feel in control	6.2%	28.4%	33.5%	31.8%	7.2%	28.7%	36.3%	27.9%	
% who said the prompt was annoying	46.3%	35.6%	11.5%	6.6%	41.7%	36.3%	14.0%	7.9%	*
% who said the prompt was something they'd expect to see	20.7%	29.2%	27.3%	22.8%	23.5%	30.4%	24.1%	22.0%	+
% who said the prompt was a useful reminder	8.2%	28.4%	31.0%	32.3%	9.4%	30.3%	29.8%	30.5%	
	Pre-engagement				**Post-engagement**				
Outcome	**Not at all**	**A little**	**Moderately**	**Very much**	**Not at all**	**A little**	**Moderately**	**Very much**	***p***
% who said the prompt was easy to understand	6.0%	20.0%	29.9%	44.1%	4.2%	19.2%	28.0%	48.6%	+
% who said the prompt made them feel in control	7.45%	30.0%	35.3%	27.2%	6.0%	27.2%	34.5%	32.3%	**
% who said the prompt was annoying	40.5%	38.1%	13.2%	8.2%	47.3%	33.9%	12.4%	6.4%	**
% who said the prompt was something they'd expect to see	21.3%	29.6%	25.5%	23.6%	22.8%	30.0%	25.9%	21.2%	+

% who said the prompt was a useful reminder	10.7%	29.1%	30.8%	29.4%	7.1%	29.6%	30.0%	33.3%	*

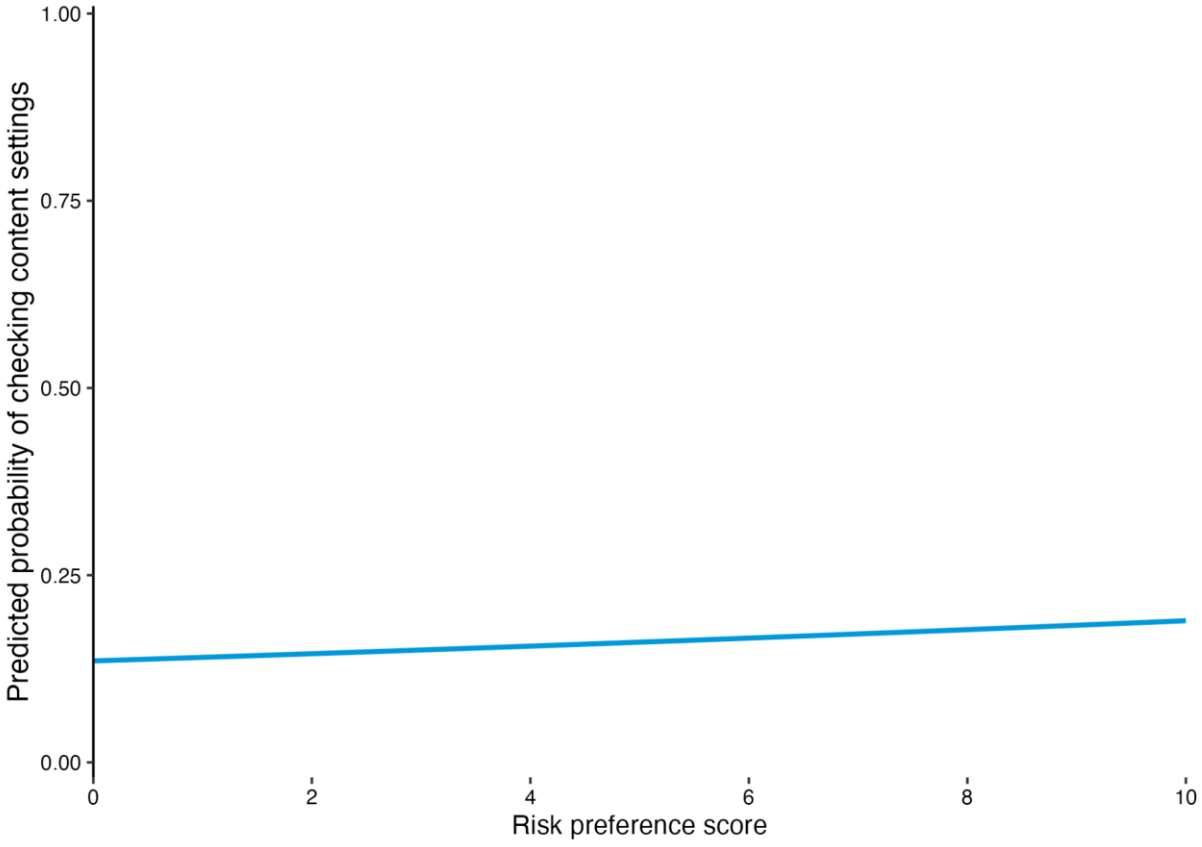
** p < .01, * p < .05, + < .1
This table reports the results of two regressions for each outcome; one regression comparing Empowerment vs. Process messages and one regression comparing Pre-engagement vs. Post-engagement timings.
Regressions exclude the control arm and control for age, gender, income, ethnicity, platform use, risk preference, mood, and energy level.
Significance is not corrected for multiple comparisons.

4.4.7 Primary outcome by psychological variables

Risk preference

A higher risk preference on social media was significantly associated with being more likely to check their settings (β = 0.046, Odds ratio = 1.047, $p < .05$). Results are shown in Figure 13.

Figure 13: Predicted probabilities of checking content settings by risk preference score.



Mood before experiment

There was no significant relationship between whether participants checked their settings and their mood before the experiment (β = 0.004, Odds ratio = 1.004, $p > .05$).

Energy level before experiment

There was no significant relationship between whether participants checked their settings and their energy level before the experiment (β = -0.005, Odds ratio = 0.995, $p > .05$).

4.4.8 Primary and secondary analysis excluding those who saw the prompt at the end of the feed

Of those in the Post-engagement arms, 1,270 participants saw the prompt after disliking a sensitive post and 175 saw the prompt at the end of the feed. Because there is not a rough split between these groups, we reran the primary and secondary analysis excluding the minority group of participants who saw the prompt at the end of the feed. This was done as a sensitivity check to make sure those who saw the prompt at the end of the feed were not affecting our results as they may have had different motivations to check their settings. As this analysis is exploratory, we do not correct for multiple comparisons.

Primary analysis

Results of the primary analysis excluding participants who saw the prompt at the end of the feed were consistent with the main primary analysis. Participants who saw the Empowerment or Process messages were significantly more likely to check their settings than those in the Control arm, who didn't see any prompt (16.8%, $p < .01$ and 23.0%, $p < .01$ respectively, compared to 4.3% in the Control arm). Participants who saw the Process message were significantly more likely to check their settings than those who saw the Empowerment message, $p < .01$.

Participants who saw the prompt before or after engaging with content were significantly more likely to check their settings than those in the Control arm, who didn't see any prompt (17.7%, $p < .01$ and 22.4%, $p < .01$ respectively, compared to 4.3% in the Control arm). Participants who saw the prompt after disliking a sensitive post were significantly more likely to check their settings than those who saw it before engaging with content, $p < .01$.

Post-hoc, we checked descriptively the proportion of participants who checked their settings in the Post-engagement arm when they saw the prompt after disliking a sensitive post and those who saw it after the last sensitive post. 22% checked their settings after seeing the prompt after disliking a sensitive post and 15% did so after seeing the prompt after all sensitive posts. This was not tested for significant differences due to the small number of participants who saw the prompt after all sensitive posts.

Secondary analysis 1

Results of the primary analysis excluding participants who saw the prompt at the end of the feed were consistent with the main secondary analysis 1. Participants who saw the Process message after disliking a sensitive post were significantly more likely to check their settings than those who saw the same message before engaging with content, $p < .01$ (26.7% vs. 19.6% respectively). Participants who saw the Process message after engaging with content were significantly more likely to check their content settings than those who saw the Empowerment message at the same time, $p < .01$ (26.7% vs. 17.8% respectively). There were no significant differences between the Empowerment Pre-engagement and Empowerment Post-engagement arms.

Secondary analysis 2

Results of the primary analysis excluding participants who saw the prompt at the end of the feed were consistent with the main secondary analysis 2. There were no significant differences in whether the final content settings matched their preferences between the Empowerment and Process messages,

or between participants who saw the prompt before engaging with the feed and those who saw it after disliking a sensitive post.

4.5 Exploratory Descriptives

4.5.1 Gear icon

Overall, 11% participants clicked the gear icon during the experiment. Of those who did, 65% checked their settings through the gear (7% of the full sample).

In the treatment arms, 15% checked their setting through the prompt and 8% checked through the gear icon (7% after seeing the prompt).

4.5.2 Clickthroughs

6 participants clicked to learn more about sensitive content in the prompt asking them if they wanted to check their settings. 1 participant clicked to learn more about sensitive content when reviewing their settings.

4.5.3 Decision to check

In the follow-up survey, participants were asked why they checked their content settings when browsing through WeConnect.

The top reasons participants chose to check their content settings (n = 587) was that they saw content that annoyed them (36%) and that they were curious to see what would change (34%). For the full list of response results, see Table 7.

Table 7. Why participants checked their content settings.

Why did you choose to check your content settings? (Participants could select more than one option, n = 587)	
I saw content that annoyed me	36%
I was curious to see what would change	34%
I wanted to see what my settings currently looked like	28%
I saw content that upset me	28%
I didn't like the content on WeConnect	17%
I didn't mean to check my content settings	7%
Other	4%

Participants who selected "Other" were able to provide a free text response. In the free text responses, participants also said they checked their settings because they didn't like the content they saw ("The algorithm wasn't set up for me yet and there were no friends to connect with, so it was going for polarised content to see which end of the scale I was at. But I was shocked it went straight for right wing hate speech as default, opposite to my beliefs/values. I wouldn't go onto this platform unless a

lot of friends were using it.") or because the prompt prompted them to ("It came up for me and not the other way around", "Had a prompt asking if I wanted to change settings, so I did").

For those who didn't check their content settings, the main reasons for not doing so were that they were curious to see what content is available on the platform (41%), they didn't know they could (30%) and they liked the content they saw on WeConnect (23%). For the full list of response results see Table 8.

Table 8. Why participants did not check their content settings.

Why didn't you check your content settings? (Participants could select more than one option, n = 3,015)	
I was curious to see what content is available on the platform	41%
I didn't know I could check my content settings	30%
I liked the content I saw on WeConnect	23%
I don't care about my content settings	17%
I don't understand what content settings are	13%
Other	3%

Participants who selected "Other" were able to provide a free text response. In the free text responses, some participants also said that they didn't check their settings because they didn't want to be distracted from the feed ("Too busy looking at other posts", "Thought it would take me off the page"), they wanted to see all the content available before they make a choice ("I wanted to know what was posted on the website without any setting changes before I put any on to see if this is a site I'd want to join, most sites also take some content choices from set up") or didn't realise they could ("I didn't even consider that there might be other settings", "Couldn't see where to check them").

4.5.4 Sentiment to the Control arm

In the follow-up survey, participants in the Control arm were reminded that they could change their content settings through the gear icon on WeConnect and asked a few questions about it. 72% of participants said this was moderately or very much something they would expect to be able to do on a social media platform, but only 36% said it was clear how they could do this. 43% said they felt in control of the content they saw on WeConnect.

4.5.5 Previous experience with content settings

The most popular way participants have previously controlled the content they see on social media was by unfollowing or blocking a user (59% and 55% respectively) or by clicking a button to state their preferences. 22% said they have previously changed settings on how much sensitive content they want to see. Full results are in Table 9. Participants who selected "Other" were able to provide a free text response. In the free text responses, some participants also said they had "Unfriended or left groups", "Made my profile private", and "Left social media alone for a while".

Table 9. How people have previously controlled the content they see on social media.

What, if anything, have you done previously to control the content you see on social media? (Participants could select more than one option)	
Unfollowed a user to not see what they're posting	59%
Blocked a user to not see what they're posting	55%
Clicked a button to state your preference (e.g. 'Not interested' or 'See less like this') on a post	52%
Reported a post	42%
Blocked certain types of content from appearing on my feed by using keywords	33%
Changed my settings on how much sensitive content I'd like to see (i.e. content that doesn't go against the platform's Community Guidelines, but refers to topics some people don't want to see, such as violence, hate speech and misinformation)	22%
Nothing (exclusive)	11%
Other	< 1%

5. Summary and Limitations

5.1 Summary

The focus of this trial was to see if any of the interventions would motivate people to make an active choice regarding their settings. We sought to measure this through whether people clicked through to review their settings either when prompted (in the intervention arms) or through accessing a gear icon while on a simulated social media feed. We tested whether varying the timing (Pre- and Post-engagement with the feed) and the framing (Empowerment or Process focused) of the prompt encouraged people to check their settings (see Table 10).

Table 10: Overview of intervention arms in trial

	Empowerment message "Your feed, your choice – you can choose the amount of sensitive content that you see."	**Process message** "It takes just two steps to check and update your content settings."
Pre-engagement Prompt appears after one post in their feed	Arm 1: Pre-engagement & Empowerment	Arm 2: Pre-engagement & Process
Post-engagement Prompt appears after disliking a sensitive post or after scrolling through all sensitive posts in feed	Arm 3: Post-engagement & Empowerment	Arm 4: Post-engagement & Process

It was theorised that the Pre-engagement prompts would be timed for when participants are in a 'cold' state whereas the Post-engagement prompts would be timed for when participants are more likely in a 'hot' state having seen some sensitive content on the platform.

Prompting people was an effective way to motivate them to make an active choice. Overall, participants who viewed a prompt message were significantly more likely to check their content settings than those who did not receive a prompt (16.7% for the Empowerment message and 22.5% for the Process message, compared to 4.3% for the Control message).

A message highlighting how quick the process motivated participants more than a message letting them know they were in control of their feed. The Process message, which highlighted that it only took two steps to review their settings, performed better when it came to motivating participants to review their settings.

Prompts delivered after people had viewed a social media feed were more likely to motivate people to review their settings. Participants who saw the prompt after engaging with the feed were significantly more likely to check their settings than those who saw the prompt prior to seeing any content (21.5% for the Post-engagement prompts compared to 17.7% for the Pre-engagement prompts, compared to 4.3% for the Control arm). In comparing the individual arms to each other, participants who saw the Process message after engaging with the feed were most likely to review

their settings (25.2% of users in this arm reviewed their settings, compared to 4.3% in the Control arm, and between 15.9%-19.6% in the other three intervention arms).

Some exploratory evidence suggests that prompts increased the alignment between participants' content settings and their stated preferences, but the timing and framing of the prompt did not matter. Overall, the final content settings matched people's preferences in 44.1% of the participants who saw a prompt. There was no statistically significant difference between this alignment in the Pre- and Post-engagement arms, and no differences between the two types of messages. An exploratory, post-hoc analysis showed that the intervention arms led to greater alignment with user preferences than the Control arm (36.1% alignment with preference in the Control arm, compared with 44.1% in the intervention arms).

People who saw the prompt after engaging with the feed were significantly more likely to correctly recall that they could change their content settings than those who saw the prompt before engaging with the feed. Participants in the Post-engagement arms also spent longer viewing the prompt to review their settings, when compared to the Pre-engagement arms. There were no significant differences between the two types of messages on these measures. When it came to time spent reviewing their settings once they had clicked through, participants who saw the Process message spent longer on this than those who saw the Empowerment message.

Overall, sentiment measures were similar across the trial arms in terms of perceived ease of understanding, feeling of control and usefulness. The Process message was seen as more annoying, but it did not backfire in terms of people choosing to review their settings. Participants who saw the Empowerment message were more likely to say this was something they would expect to see on social media compared to those who saw the Process message. Nevertheless, these sentiments did not appear to be a driving mechanism of behaviour.

We did not find differences in the reported mood or energy levels before the experiment. At the same time, a higher risk preference on social media was associated with being significantly more likely to check their settings.

Taken together, these findings suggest that prompting people to review their content settings after they have spent time engaging with a feed motivated people to do so, particularly when people were prompted with a message highlighting how easy it was for them to do so.

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. 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. Finally, this trial included a training task which primed participants to interact with the content and increased the engagement levels compared to similar trials without the training task. Despite these limitations, we believe online RCTs are a useful tool for building the evidence base.

6. Annex

Annex A: 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 11). Results from these regressions should be interpreted with caution.

Table 11. The results of the Brant test on the proportional odds assumption for each sentiment variable.

Exploratory outcome 6: Ease of understanding			
Message comparison	**χ^2**	**df**	***p***
Omnibus	25	30	.72
process	2.29	2	.32
Timing comparison	**χ^2**	**df**	***p***
Omnibus	25.06	30	.72
Post-engagement	2.39	2	.30
Exploratory outcome 7: Feeling of control			
Message comparison	**χ^2**	**df**	***p***
Omnibus	41.79	30	.07
process	3.26	2	.20
Timing comparison	**χ^2**	**df**	***p***
Omnibus	38.79	30	.13
Post-engagement	0.43	2	.81
Exploratory outcome 8: Annoying			
Message comparison	**χ^2**	**df**	***p***
Omnibus	46.67	30	.03
process	0.43	2	.81
Timing comparison	**χ^2**	**df**	***p***
Omnibus	48.74	30	.02
Post-engagement	2.35	2	.31
Exploratory outcome 9: Something they'd expect to see			
Message comparison	**χ^2**	**df**	***p***
Omnibus	45.91	30	.03

process	2.23	2	.33
Timing comparison	**χ^2**	**df**	***p***
Omnibus	44.88	30	.04
Post-engagement	1.05	2	.59
Exploratory outcome 10: Useful			
Message comparison	**χ^2**	**df**	***p***
Omnibus	48.13	30	.02
process	0.49	2	.78
Timing comparison	**χ^2**	**df**	***p***
Omnibus	54.43	30	.00
Post-engagement	6.76	2	.03

Annex B: User journey

Figure 14: The training feed instructions.

Figure 15: The training feed. Participants see 3 non-sensitive posts and are asked to like at least one post, dislike at least one post, and repost at least one post to progress through the survey.

Figure 16: The main feed instructions.

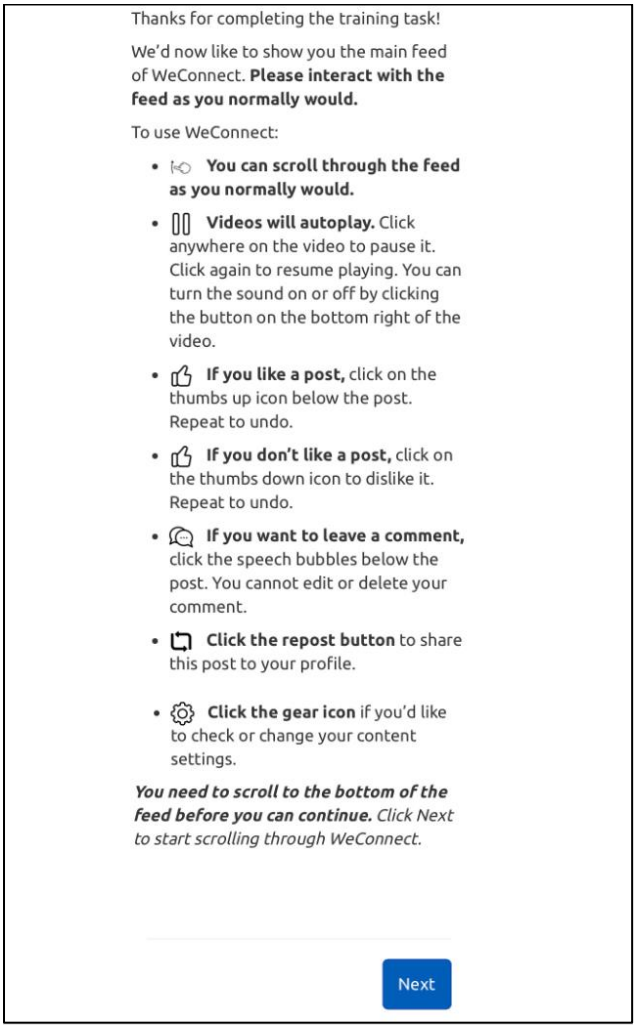


Figure 17: The main feed.

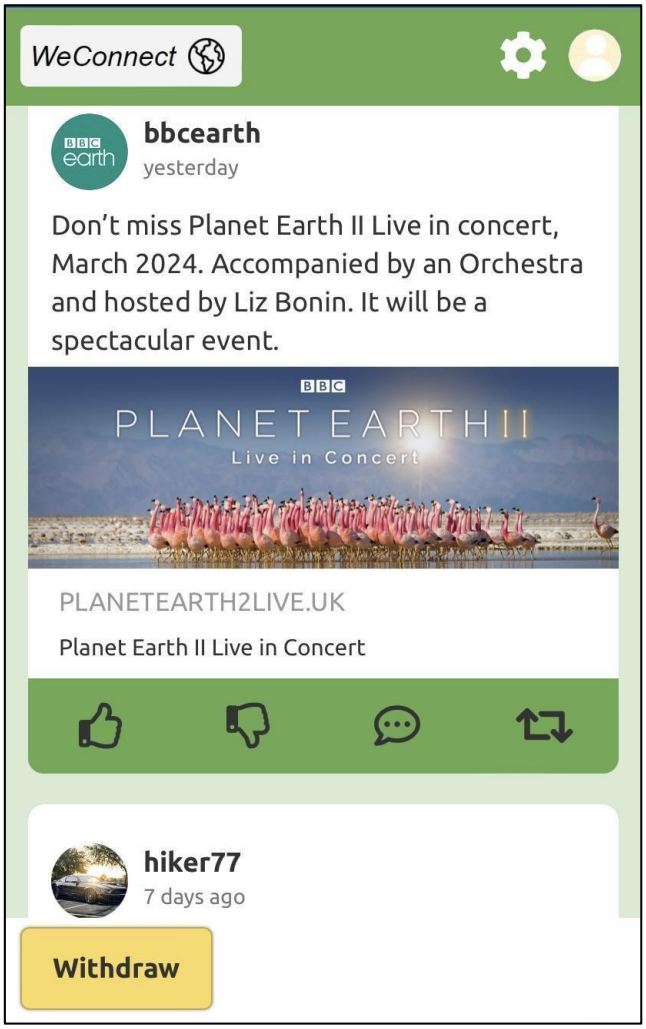


Figure 18: The main feed after a participant clicks on the gear icon.

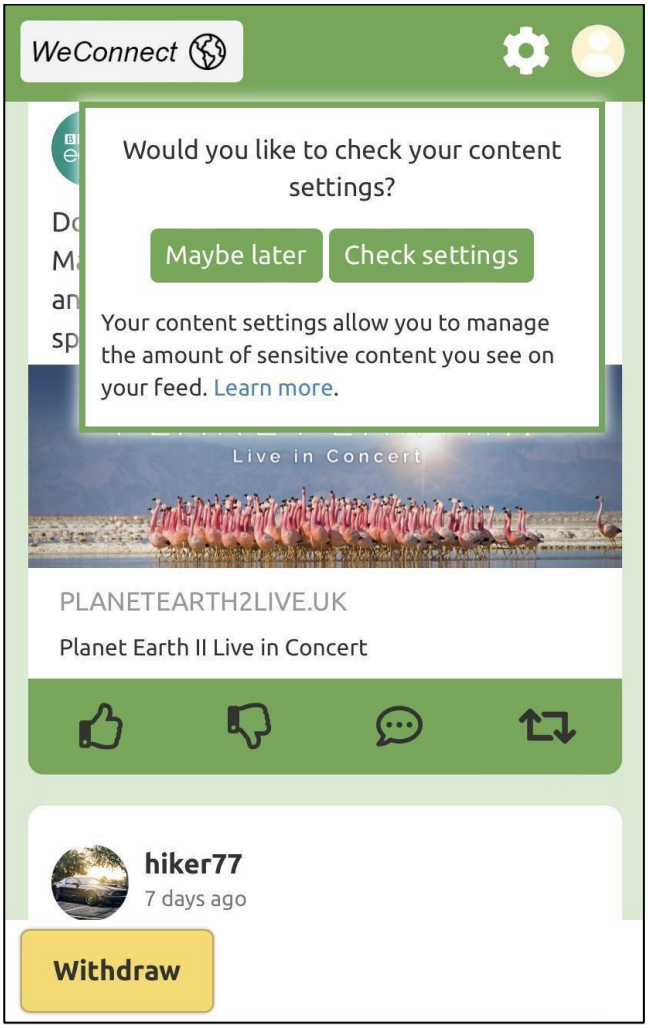


Figure 19: The left screen shows the review prompt for participants already viewing all content types (and therefore shown to participants the first time they click to check their settings). The right screen shows the review prompt for participants who already checked their settings and chose to view reduced sensitive content.

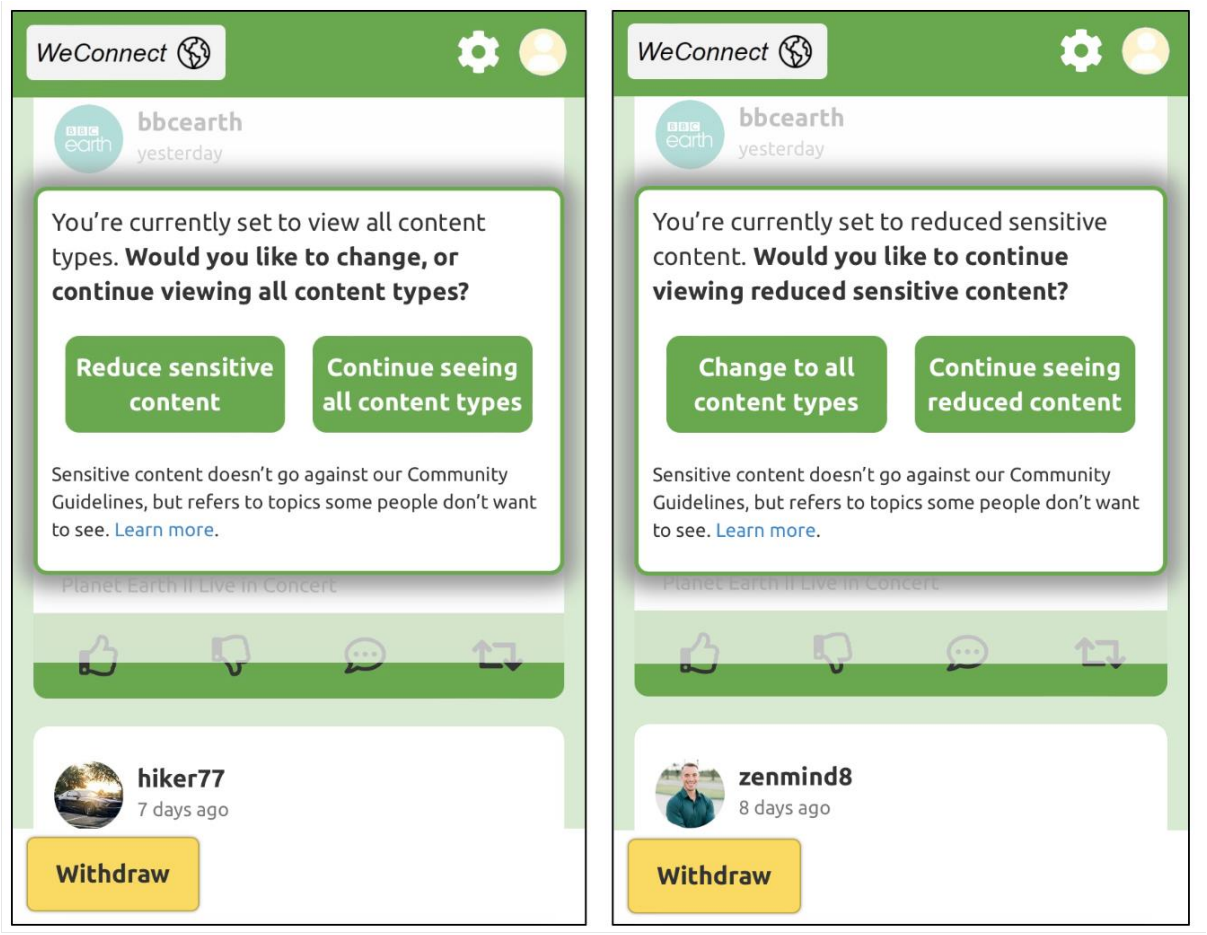


Figure 20: The main feed after a participant clicks the "Learn more" button through the gear icon or the review prompt.

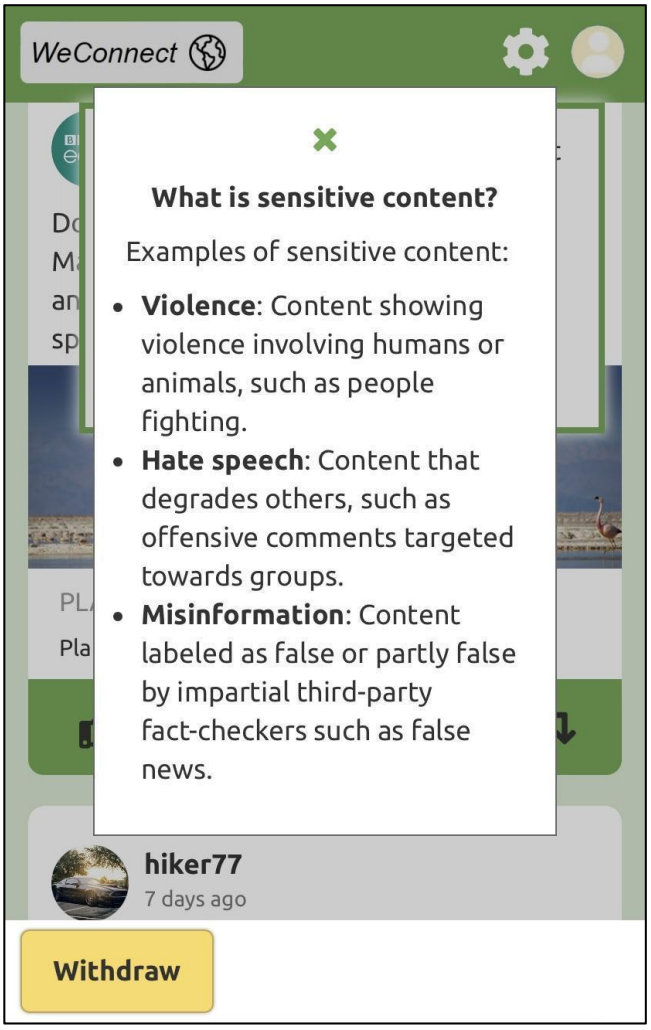


Figure 21: The main feed after a participant changes their settings. The left screen shows a feed for a participant who has chosen to see all content types. The right screen shows a feed for a participant who has chosen to see reduced sensitive content.