30 October 2023

User Attitudes towards On-Platform Interventions

Qualitative findings

YouGov Qualitative

YouGov®

Living Consumer Intelligence | business.yougov.com

Table of contents

YouGov®

1. Background and Methodology

YouGov®

Research objectives

Ofcom commissioned YouGov to conduct qualitative research to gauge online users' **understanding and perceptions of on-platform interventions,** focusing on 5 broad types (labels, overlays, notifications, resources and prompts) following a sampling survey.

Ofcom's key aim was to gain insights on **how to improve or change on-platform interventions,** based on what users think should be considered when developing and designing them. This insight was to inform part of Ofcom's Making Sense of Media programme, developing best practice design principles for platforms.

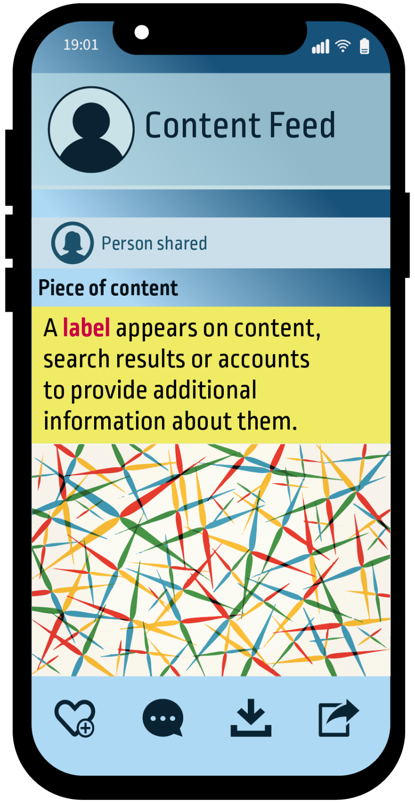
To address these aims, YouGov's qualitative study explored:

- The perceptions of users who have / have not experienced interventions;
- The users' attitudes and emotions when encountering an intervention;
- To what extent users can understand or recognise the distinctions between interventions and advertisements/cookies on websites.

Introduction to Interventions

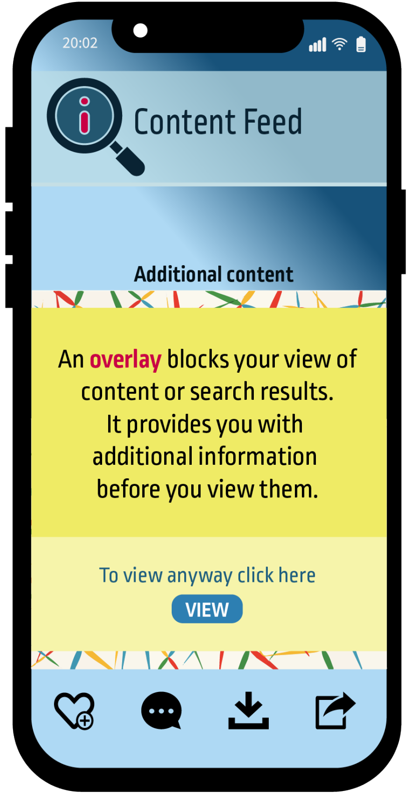
‘Online interventions’ are ways that social media platforms, gaming platforms and search engines provide information to help their users while they are online e.g., by providing real-time information on the content they are about to view/engage with. Interventions are provided by the platforms themselves, not users of the platforms. The following five interventions were covered in the research:

Labels



A **label** appears on content, search results or accounts to provide additional information about them.

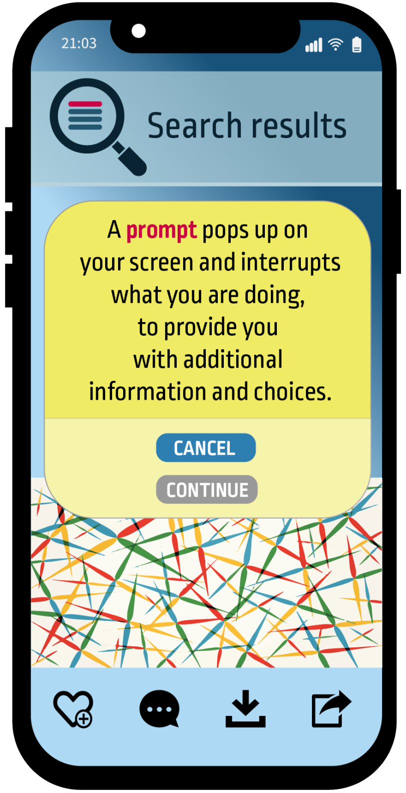
Overlays



An **overlay** blocks your view of content or search results. It provides you with additional information before you view them.

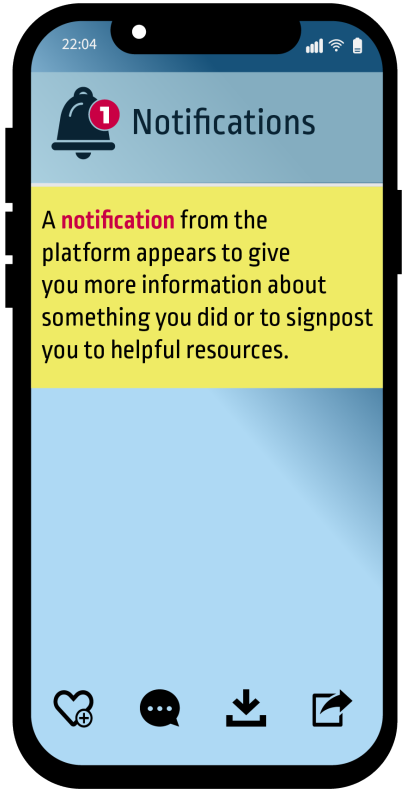
To view anyway click here

Prompts



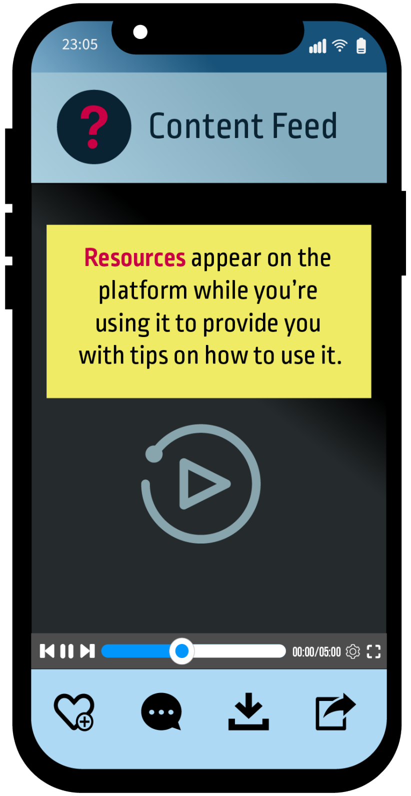
A **prompt** pops up on your screen and interrupts what you are doing, to provide you with additional information and choices.

Notifications



A **notification** from the platform appears to give you more information about something you did or to signpost you to helpful resources.

Resources



Resources appear on the platform while you’re using it to provide you with tips on how to use it.

Sampling Survey - Summary

1072 people were surveyed by YouGov's custom quantitative team in November and December 2022. Respondents were from YouGov's online research panel. The sample was representative of UK internet users aged 13-84.

Key findings:

1. Participants reported encountering all types of intervention with similar frequencies, apart from resources which was particularly low. Prompts were the most encountered intervention type but only by a small margin.
2. Younger people were more likely to have seen any type of intervention. This was most likely due to younger people being on social media more frequently than older people.
3. The platforms where interventions were seen the most were YouTube, Instagram and TikTok; this may be because these platforms are used more by younger generations. Age differences were also apparent in people's attitudes towards platforms. Younger people thought they were more aware of the tools that the platforms provided and were confident using them.
4. There was also a tendency among younger people and those with mental health conditions to be a bit more impulsive in their activities online. These groups were more likely to share or like things that they did not fully read or watch, as well as post, view, or search for things they later regretted.

Qualitative methodology

Online Diary

- **2x two-week online diary with two target audiences:**
  - Adults 18+ (37 participants)
  - Teenagers 13 – 17 (12 participants)
- The sample included a mix of participants who had & had not encountered interventions
- The diary included a mix of ages, gender, ethnicity, socio-economic backgrounds and locations
- Fieldwork was conducted between Tuesday, 14th February and Monday, 27th February 2023

Pause and Reflect

- The discussion guide and stimulus for the text-based focus groups were revised based on observations and findings from the online diary stage

Text-based Focus Groups

- **8x text-based focus groups** split between adults and teenagers, 4 groups with adults (G1, G2, G3, G4)
  - 2 groups with parents of teenagers aged between 13 to 17 (G5, G6)
  - 1 group with teenagers aged 13-15 (G7)
  - 1 group with teenagers aged 16-17 (G8)
- Each group had 8 – 10 respondents in it and lasted 90 minutes
- The focus groups included a mix of ages, gender, ethnicity, socio-economic backgrounds and location
- Fieldwork was conducted between 22nd March and 28th March 2023

2. Key Findings from the qualitative study

Interventions were appreciated as a warning and a source of information

- Participants **disliked** the perceived **intrusiveness** of interventions. The majority felt annoyed at the interruption or became apathetic towards interventions, choosing to ignore them, which increased with exposure over time.

- However, both adults and teenagers felt that interventions were **useful - mostly for others rather than themselves - for flagging sensitive or upsetting content**. Adults in particular thought that interventions could be **valuable sources of information** when online for more vulnerable audiences (e.g., children or the elderly).

- **Notifications and Prompts** intervention types were the **hardest to recognise** for all participants regardless of age. They were most often confused with standard notifications (e.g., email alerts, social media likes) and with pop-ups (e.g., cookie selections, ads).

- Participants reflected that those who are not as digitally literate and young children may find interventions hard to understand.

- There was evidence of **behavioural change**, as some participants across all ages reported having avoided posting or viewing content following an intervention. There were however **paradoxical effects** among adults: some reported having learnt to reword their online content to avoid being sent an intervention; others claimed to have stopped using platforms whose interventions were 'too disruptive'.

Parents and teenagers had different views on the utility and behavioural effects of interventions

- **Parents felt interventions were helpful for younger teenagers,** who need guidance and protection online, but may work less well for 16+ year-olds who would be able to make more informed decisions online.

- **Younger teenagers (13-15) reported feeling annoyed** at the interruption caused by interventions and tended to ignore them. They acknowledged that interventions can be helpful when informing them about privacy or breaching guidelines, as it made them feel the platforms cared about their online safety.

- **Older teenagers (16-17)** saw interventions as helpful in the case of sensitive content but also irritating, especially when they block content inaccurately. This audience was **generally sceptical** of how content is flagged, feeling that the process is unclear and not transparent.

- **Parents were worried** interventions may push children to bypass them and access content **out of curiosity.**

- However, **younger teenagers** argued that they **have never accessed content blurred by overlays** because the intervention made clear it was unsafe. Whereas **older teenagers** said they had **often proceeded to view** the content to check on the platforms' accuracy in flagging harmful content

How participants experience interventions

- **The internet is filled with interruptions vying for users' attention**. It is crowded with advertisements, calls to action (i.e., ‘please subscribe’ prompts) and cookie/privacy setting notifications.
- **These interruptions can create frustrations**, as the participants find them hard to distinguish from adverts, as they use the same mechanisms and are crowded out by the frequency of the adverts. **This makes it harder for interventions to cut through and make an impression.**
- Interventions **are not seen to have a great influence on behaviour** (as stated by most participants and observed in the diary study) but do act as an awareness point to help users navigate the internet.
- **Most teenagers believe interventions are helpful and offer protection for them online, especially for harmful content** and, to a lesser extent, influencing their behaviour (e.g., to limit screen time). A few admitted it may make them curious to view the content in question, but it is not possible to compare the benefits to others vs the risks of individual behaviours within this methodology.

Perceived value of interventions

- **Participants felt there is clear value in the use of interventions despite them being frustrating at times**. For instance, they are valued as a tool to stop users being able to view harmful content instantly, and also to raise awareness to users of their own potentially harmful behaviour online.

- **Interventions are most needed on user-to user services (i.e. social media).** This is because participants feel there is a higher perceived risk of harmful content being posted or users acting inappropriately.

- **Although most adults feel they would ignore interventions, they believe they are vital for children** to help reduce exposure to possible harms from being online.

- Interventions relating to illegal content could be used to offer support for those who have seen the content and as initial warning of behaviour before escalation. However, the participants felt **due to the ability to ignore interventions and the seriousness of illegal content, further work to prevent illegal content appearing online altogether is necessary.**

- Participants feel **using interventions to offer support and advice around harmful content (i.e., self-harm) should not be solely from the platform**, as they are not seen as professionals in these areas. Rather, it should be done with respected and trusted sources i.e., charities and health organisations.

Considerations for intervention improvement

- **Participants want interventions to be used as both a warning system and a tool to help support and educate users online**, particularly to inform them why an intervention was necessary.
- However, there was a **lot of debate especially around interventions relating to misinformation** and who gets to define it. This is because individuals may disagree on what information is contentious/harmful.
- **For adult users, interventions could be made optional**. Adults felt they should have the freedom to choose whether interventions can prevent them seeing harmful content or contentious information/opinions.

3. Diary Findings

Participants submitted two online diaries: One tracking their online behaviour and a second recording if/when they encountered interventions.

The purpose of the stage was to capture their real-life activity and understand their reflections across all five interventions.

Users encountered several interventions per week, and found overlays and labels the most useful

Effectiveness of Interventions

- Participants mostly came across interventions on social media, including TikTok, Facebook, Instagram and Twitter, around **1 or 2 times a week**. The most viewed interventions were **overlays, labels and prompts**.

- Our analysis showed that not all interventions were considered equally effective, and some were perceived to be more useful than others:

  - **Overlays and labels about misinformation** were considered the most useful by the participants who encountered them, as they appreciated being warned about potentially upsetting or misleading content. Some participants thought that these interventions showed that the platform 'cared' about its users and was trying to provide them with a good user experience.

  - **Labels about paid promotions** were also considered useful by those who saw them on social media, because they offered transparency and helped participants make informed decisions knowing that the information presented was not impartial.

  - **Overlays about sensitive topics** were met with mixed feelings: while some participants wanted to know in advance (to avoid viewing the content), others questioned on what basis content was being flagged when for them it was not 'sensitive at all'.

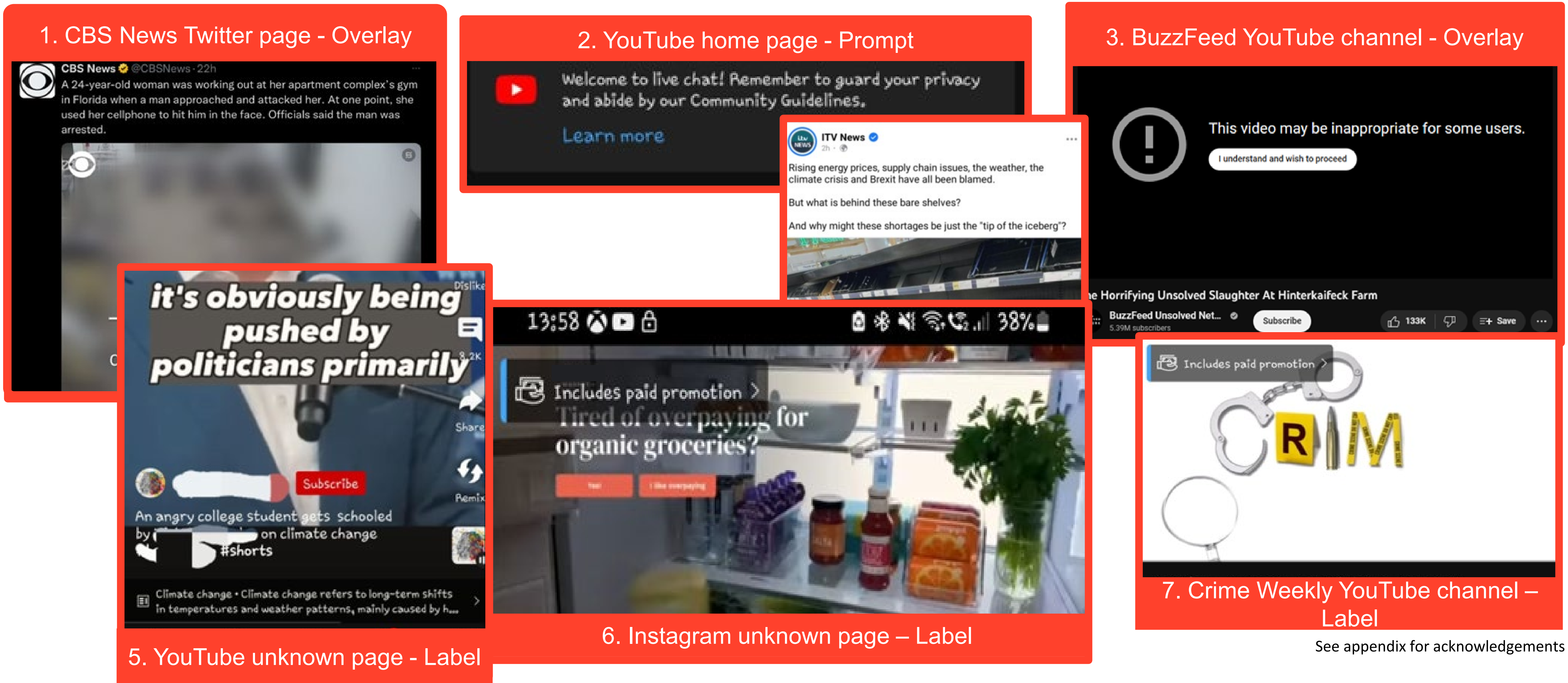
- It appears that the **utility of interventions** changed over time. For instance, some participants noted that **prompts** to limit their time on a platform felt **more effective early on**, however their **repetition** meant they got used to them and did not engage anymore.

"I like [the overlay] intervention. It shows the social media app or website is monitoring what is being uploaded and viewed".
18+ Male, Online diary

"[Overlay is] a good intervention as it can prevent you from seeing something you don't want to. Let's you decide if you want to view the video or image etc. rather than see it without notice and wish you hadn't seen it."
18+, Female, Online diary

"I've seen [prompts] on TikTok and it told me that the post I was looking through was reported as fake information. This made me think that the outlet was fake so I stopped looking into it and moved on."
13-17, Male, Online diary

Some examples of participants' own encounters of interventions



1. CBS News Twitter page - Overlay

CBS News @CBSNews · 22h
A 24-year-old woman was working out at her apartment complex's gym in Florida when a man approached and attacked her. At one point, she used her cellphone to hit him in the face. Officials said the man was arrested.

2. YouTube home page - Prompt

Welcome to live chat! Remember to guard your privacy and abide by our Community Guidelines.
Learn more

3. BuzzFeed YouTube channel - Overlay

This video may be inappropriate for some users.
I understand and wish to proceed

ITV News
2h

Rising energy prices, supply chain issues, the weather, the climate crisis and Brexit have all been blamed.

But what is behind these bare shelves?

And why might these shortages be just the "tip of the iceberg"?

5. YouTube unknown page - Label

An angry college student gets schooled by [redacted] on climate change #shorts

Climate change • Climate change refers to long-term shifts in temperatures and weather patterns, mainly caused by h...

6. Instagram unknown page – Label

Includes paid promotion

7. Crime Weekly YouTube channel – Label

Includes paid promotion

See appendix for acknowledgements

Users tended not to interact with interventions but stopped to pay attention, even if only for a matter of seconds

Interaction with Interventions

- Most participants **did not interact** with the interventions. They either ignored them or took notice of them and moved on.
- **Those who 'took notice'** of the intervention found the information **helpful** in some cases, such as when it flagged inappropriate content. In those situations, even if the participants did not interact with it, the intervention made them stop for a few moments and pay attention. Other participants **ignored** interventions altogether and kept scrolling.
- Overall, it was felt that interventions had an **impact** in terms of **raising awareness** about online content. **This was not the case in terms of behavioural change**. For example, in most cases of overlay interventions, participants went ahead and viewed the content anyway. On some occasions, the overlay counterintuitively created a sense of FOMO (Fear Of Missing Out) as participants felt more interested or curious about the 'hidden' content.
- **Teenagers (13-17)** appeared keener to **follow the instructions** provided by interventions and reported reading through and **adjusting their behaviour** accordingly in some cases, for instance to avoid viewing flagged content.

"If I encountered [a notification], I would most likely read it because they usually provide advice which will contribute to a more enjoyable user experience overall."
13-17, Female, Online diary

"The guidelines have to be on [display] so that the app can show they are doing all they can to prevent people from posting wrong things".
18+, Male, Online diary

"Yes [I see labels] - on Instagram and TikTok mainly. They were annoying and getting in the way of using the app. I just removed and ignore them."
13-17, Female, Online diary

Many participants mistook platform interventions for adverts or notifications as they used the same mechanisms

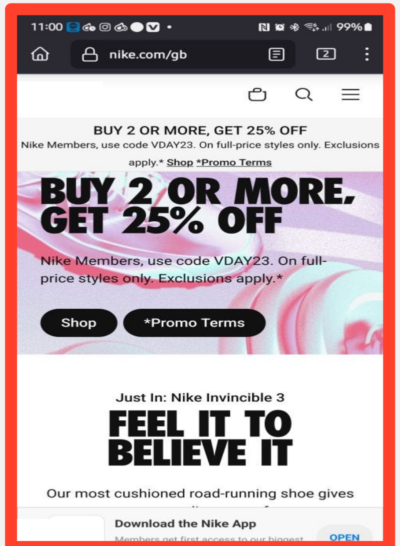
Identifying Interventions

- In terms of **understanding interventions**, participants found some interventions easier to recognise than others, in particular, overlays. However, the frequency and number of interruptions experienced by users online meant that many were unsure whether cookie selections or newsletter signups qualified as interventions.

- It was clear to participants that interventions **communicate** 'something' to them about what they are doing online, however, they were often unsure about 'who' interventions should come from (i.e., who should have the authority to deliver an intervention).

Examples of participants' uploads mistaken as interventions

See appendix for acknowledgements



BUY 2 OR MORE, GET 25% OFF
Nike Members, use code VDAY23. On full-price styles only. Exclusions apply.* Shop *Promo Terms

BUY 2 OR MORE, GET 25% OFF

Nike Members, use code VDAY23. On full-price styles only. Exclusions apply.*

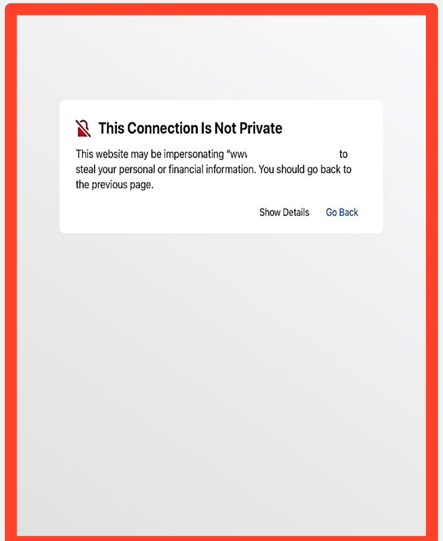
Just In: Nike Invincible 3

FEEL IT TO BELIEVE IT

Our most cushioned road-running shoe gives

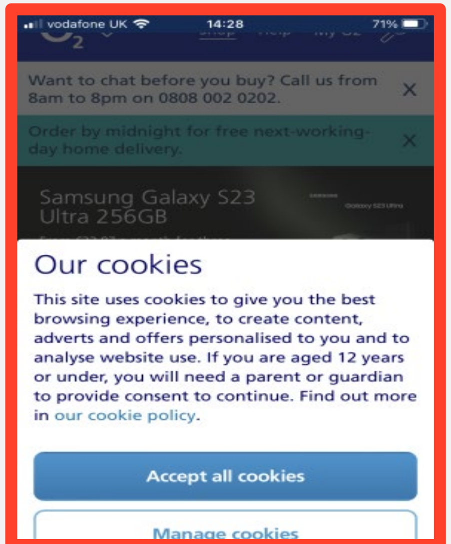
Download the Nike App

1. Promotions


This Connection Is Not Private

This website may be impersonating "www[redacted] to steal your personal or financial information. You should go back to the previous page.

2. Data encryption



Want to chat before you buy? Call us from 8am to 8pm on 0808 002 0202.

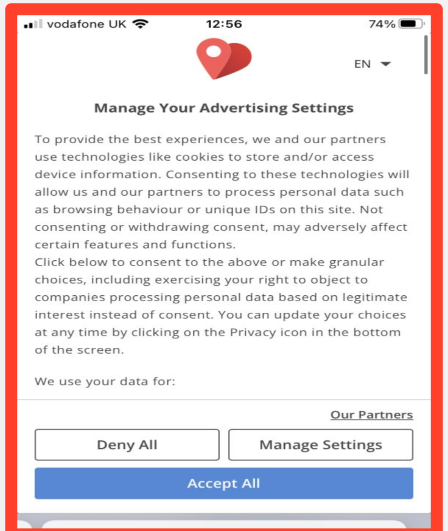
Order by midnight for free next-working-day home delivery.

Samsung Galaxy S23 Ultra 256GB

Our cookies

This site uses cookies to give you the best browsing experience, to create content, adverts and offers personalised to you and to analyse website use. If you are aged 12 years or under, you will need a parent or guardian to provide consent to continue. Find out more in our cookie policy.

3. Cookies

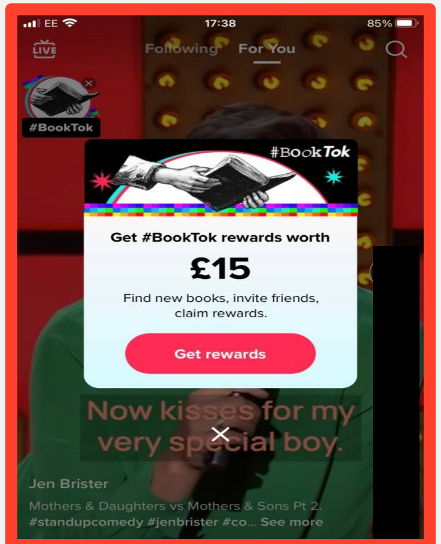


Manage Your Advertising Settings

To provide the best experiences, we and our partners use technologies like cookies to store and/or access device information. Consenting to these technologies will allow us and our partners to process personal data such as browsing behaviour or unique IDs on this site. Not consenting or withdrawing consent, may adversely affect certain features and functions.
Click below to consent to the above or make granular choices, including exercising your right to object to companies processing personal data based on legitimate interest instead of consent. You can update your choices at any time by clicking on the Privacy icon in the bottom of the screen.

We use your data for:

4. Privacy settings



Get #BookTok rewards worth

£15

Find new books, invite friends, claim rewards.

Jen Brister

Mothers & Daughters vs Mothers & Sons Pt 2. #standupcomedy #jenbrister #co... See more

[redacted]

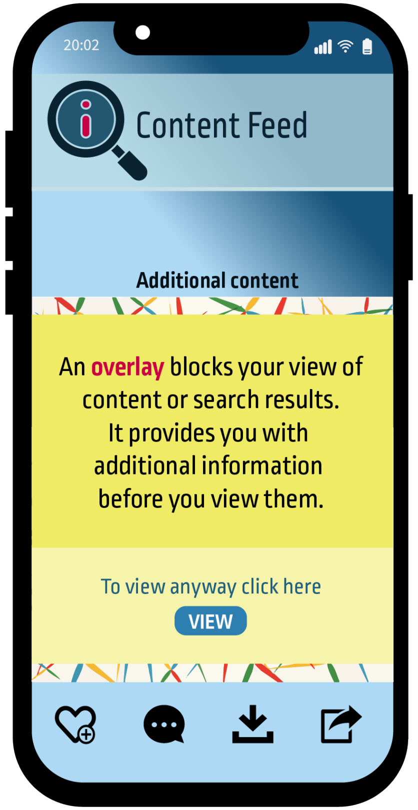
5. Ads

4. Interventions: Detailed Findings

Combined qual findings from text-groups and online diaries

1. Overlays

Definition and examples shown during qualitative stages



20:02

Content Feed

Additional content

An **overlay** blocks your view of content or search results. It provides you with additional information before you view them.

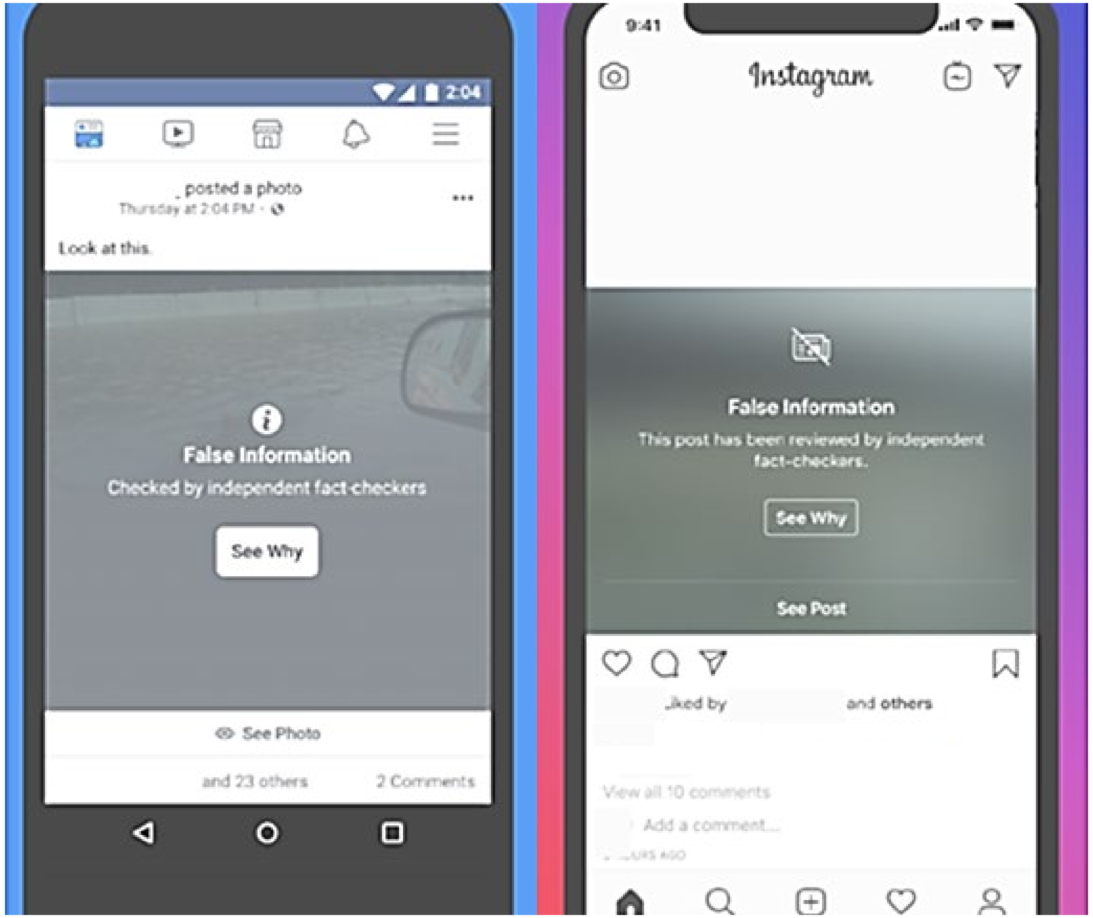
To view anyway click here

VIEW

Overlays

What is an **overlay**?

An **overlay** blocks your view of content or search results. It provides you with additional information before you view them. It may also guide you to other resources for more information or support.



See appendix for acknowledgements

Overlays

Value

Most adults and teenagers believed **overlays were useful for flagging sensitive, harmful or triggering content**. Adults thought that overlays would be **more valuable to specific audiences, such as children and the elderly** as they were more susceptible to risks online. Most age groups felt irritated by the interruption forced by overlays.

Content

Some adults and older teenagers (16-17) were **concerned about content 'policing',** questioning who decides what needs to be moderated or not. **This was related to cases of 'false' information,** whereas for sensitive content, most audiences agreed that overlays were useful and necessary, especially for children and vulnerable users.

Behaviour

Many, across age groups, reported avoiding viewing posts blurred by overlays, evidencing some behavioural change. Parents agreed that this intervention physically prevents teenagers from viewing inappropriate content, although they questioned whether teenagers would still go ahead and access the content. **Participants agreed that blocking content encouraged users to stop and think**, however they did not consider it a form of involved education.

Improvement

Although **overlays were felt to 'work' because they physically prohibit interaction with certain types of content**, adults and parents believed they could be made more visual with a clear 'stop' warning, especially for teenagers and children. Others, including teenagers, suggested the trigger type (e.g., disturbing/harmful) should be mentioned upfront and the levels of warning could be colour-coded (e.g., orange vs red).

"It's helpful because it protects people from content they may be sensitive to."
18+, Female, Online diary

"Nothing wrong with overlays the first time you fire up an app, but every time would be a pain."
74, Male, Adults focus groups

"I've experienced Instagram overlays that were very helpful, like it warned me once about a video with animal abuse, so I didn't watch it."
16-17, Female, Children focus group

"A wider banner saying 'false information' and to say what the 'false information' actually is in like a sentence below the warning. Also needs to be brighter and clearer with a bigger typeface [as an improvement]."
20, Female, Adults focus groups

Parents and teenagers agreed that overlays were useful on sensitive content, but both questioned whether it would make audiences more curious.

Parents

Parents thought that overlays were quite intrusive and could be useful only in the most serious cases of violent/inappropriate content. Generally, they felt **that overlays should explain why the content was flagged and add more context so teenagers could understand better.**

The majority **of parents feared overlays could enhance teenagers' curiosity** and push them to access the content. They thought there was not much that could be done about this.

Overall, they thought that overlays were the most successful intervention because of the interruption they provided. However, they acknowledged that flagged content can still be accessed, but this would depend on an individual's decision.

"They show that the users' actions are being watched and the next step (such as accessing flagged content) only happens with their explicit permission, so they are fully responsible for that decision."
48, Male, Parents Focus Group

Teenagers

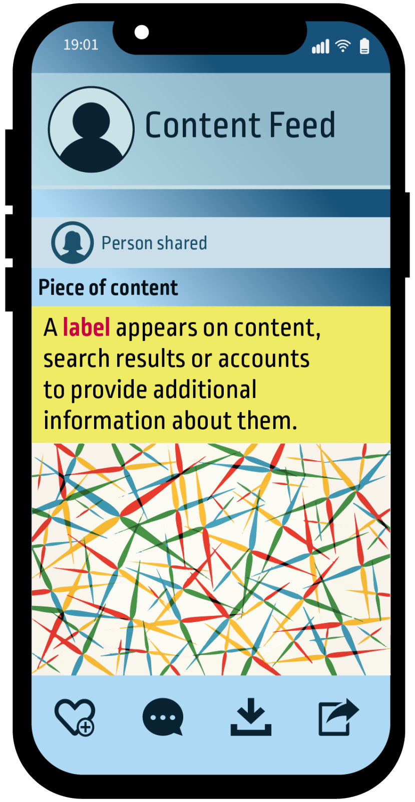
16-17-years-olds agreed **that overlays are useful as protection against distressing content but not for false information.** They suggested that in some cases they can be incorrect and hide content that was not sensitive.

They believed **that platform moderation was subjective and not objective**, and in some cases, overlays made them more curious as to access content.

This audience thought that overlays were successful in making people stop and reflect because of the physical interruption they provided, **although a minority argued that they simply 'delayed' access to content**. All agreed that the benefit of the intervention was giving users a choice as to whether to view or avoid the content.

"I think these are actually helpful for sensitive photos / videos, for example NSFW content or something potentially distressing. However, I don't think they're helpful for 'false information'."
16-17, Female, Children Focus Group

2. Labels
Definition and examples shown during qualitative stages



19:01

Content Feed

Person shared

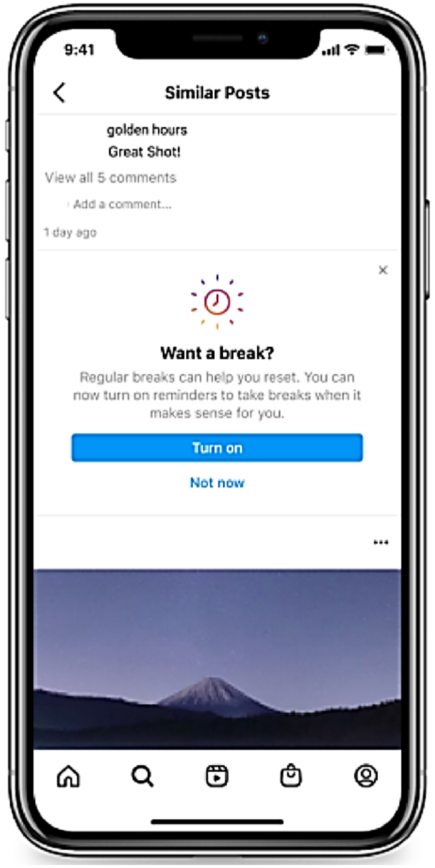
Piece of content

A label appears on content, search results or accounts to provide additional information about them.

What is a **label**?

A **label** appears on content, search results, or user accounts, to provide you with additional information about them. A label may link you to other resources for more information or support. A label does not block your view.

Labels



9:41

Similar Posts

golden hours
Great Shot!

View all 5 comments

Add a comment...

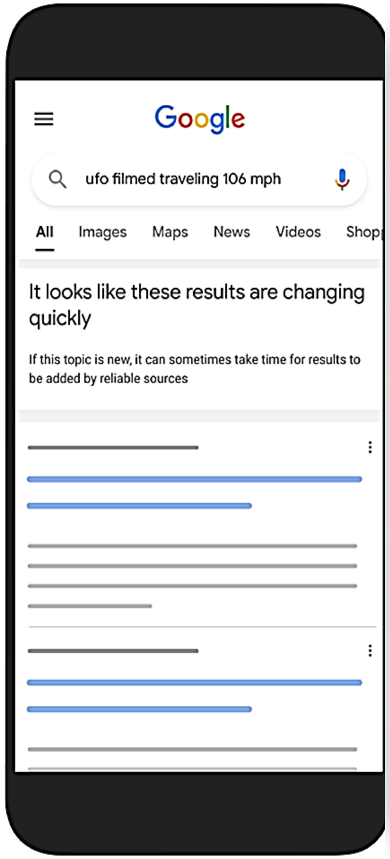
1 day ago

Want a break?

Regular breaks can help you reset. You can now turn on reminders to take breaks when it makes sense for you.

Turn on

Not now



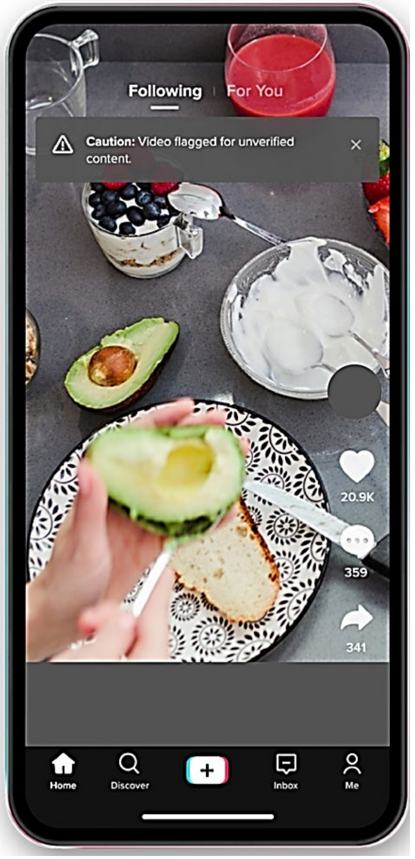
Google

ufo filmed traveling 106 mph

All Images Maps News Videos Shop

It looks like these results are changing quickly

If this topic is new, it can sometimes take time for results to be added by reliable sources



Following | For You

Caution: Video flagged for unverified content.

20.9K

359

341

Home Discover + Inbox Me

See appendix for acknowledgements

Labels

Value

Labels were seen by many **as less intrusive** than the other intervention types. They have the potential to act as **helpful reminders**. Adult participants agreed that they could be useful for those who lose track of time on social media, gaming or betting sites. Others felt patronised as they argued they were 'perfectly capable' of deciding on their own.

Behaviour

Parents felt that **teenagers may easily become 'immune' to labels** and quickly bypass them, as they are not spelling out enough the dangers and consequences of a particular action. However, **teenagers felt that labels did not impact their behaviour** not because they were not clear enough, but rather **because they were not noticeable.**

Content

Participants felt that labels were successful, in principle, in making users stop and reflect about the content they view or post. However, they argued that **their success decreases when they are not noticeable or do not stand out**. This point resonated with teenage audiences as well.

Improvement

Participants suggested that **more emphasis should be placed on 'caution'**, e.g. by using a bigger and more colourful font, or an alert symbol. Teenagers similarly felt that labels can easily go unnoticed. Adults agreed **that using brand colours* would make them stand out.** Parents added that labels should be more strongly worded, with consequences clearly spelt out, and should create engagement before the user can move on.

*colours associated with the platform where the label appears

"I think the benefits are it can help people to have more control over what they do and read online."
26, Male, Adults focus group

"The internet and social media is flooded with different opinions, views, and information looking like facts. So being pointed to the right place was good. Felt it was good that someone actually cares I get facts not someone's funny ideas."
18+, Female, Online diary

"Potential benefit [of labels] is that it warns you about the content. Drawback is that it takes responsibility away from the site/app to verify this content before presenting it to others."
46, Male, Parents focus group

"[Labels] would make me at least review my behaviour... Not so sure about my kids, as they are growing up with lots of warnings meted out around them and they tend to ignore ones they can, just to rebel against authority."
48, Male, Parents focus group

Parents and teenagers had opposing views on the 'take a break' label, but they agreed it would be useful on sensitive content.

Parents

Parents had mixed feelings on labels as they thought **they were quick to read and unobtrusive, but also easy to miss**. They felt that **labels about 'taking a break' would be useful on gaming sites and social media**, as their children could stay on them for hours. However, they recognised that, in this case too, the intervention could easily be ignored.

They thought that it could be more useful and effective on harmful content, as it would act as a warning.

Parents felt that there would be a difference in how different ages could react to labels for them, **younger teenagers would be more likely to adhere to the warnings whereas the 16+ would be likely to ignore** and feel they are 'old enough' to make up their own mind.

"I would most likely pay attention, as would my 16-year-old. I can't be sure my 12-year-old would care, and just proceed doing whatever it was anyway."

49, Male, Parents Focus Group

Teenagers

Teenagers disagreed on the 'take a break' label, as they thought it was ineffective and they would never follow it on social media or on gaming platforms.

They agreed that labels can be useful for flagging content because they signal a potential danger.

Their opinions were split on labels around search results, as they felt they knew 'reliable sources' (which they defined as well-known websites and newspapers).

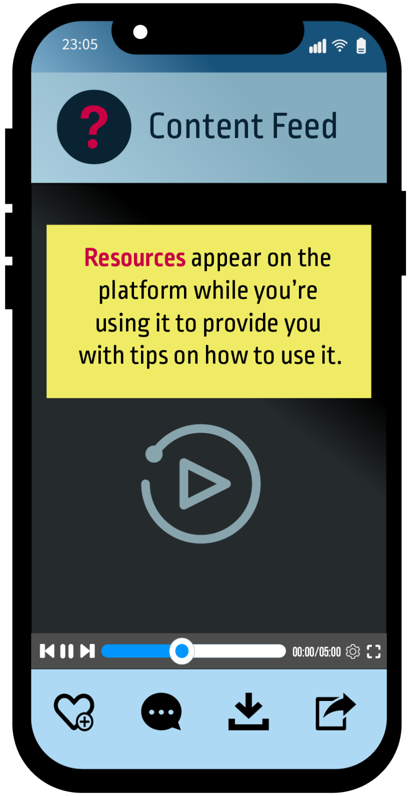
However, teenagers suggested that labels can easily be missed. **Unless made bolder, brighter and more noticeable, label interventions would not be effective** at either educating or changing online behaviours, and would rather go unnoticed.

"The colour of the label should not blend in with the rest of the app. The intervention should not be placed where the user expects junk to be, otherwise it will be ignored whether willingly or unwillingly."

13-17, Male, Online diary

3. Resources

Definition and examples shown during qualitative stages

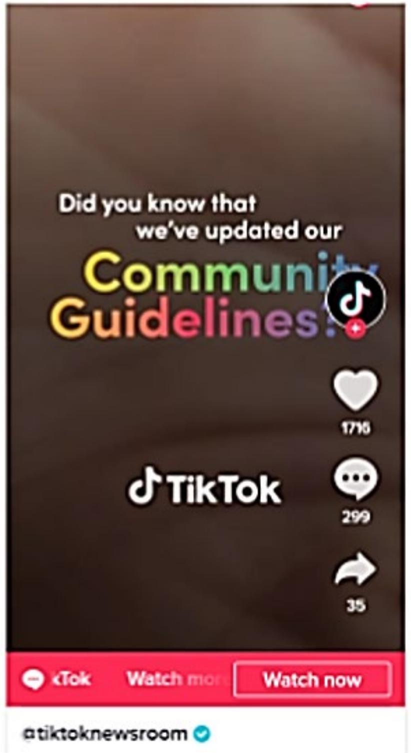


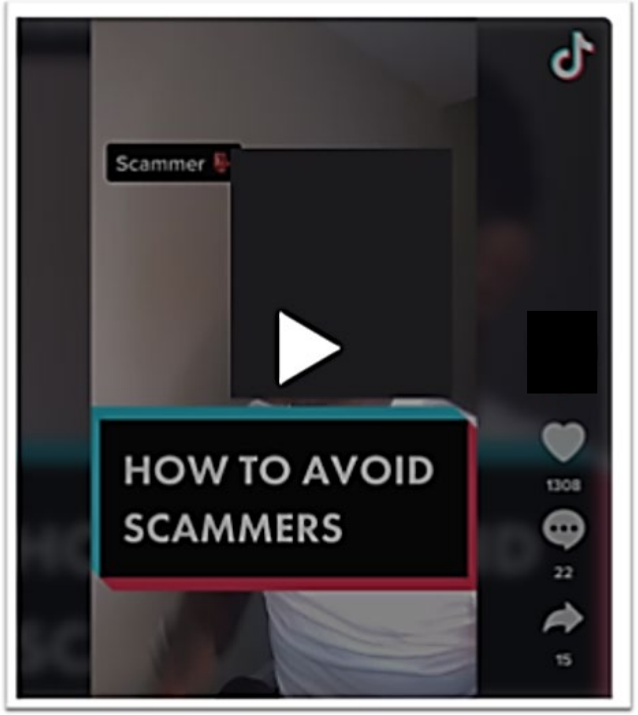
Resources appear on the platform while you're using it to provide you with tips on how to use it.

Resources

What are **resources?**

Resources are tips from the platform that appear while you are using it, such as how to report content or limit screen time.





See appendix for acknowledgements

Resources

Value

Most participants agreed that **resources could be useful to new users or vulnerable individuals such as children, older people or anyone who is not as digitally literate.** Some adults liked that they had the option whether or not they wanted to view the resource. Teenagers found them particularly entertaining and engaging, because of the video format.

Behaviour

All participants, regardless of audience type, agreed **that resources can change behaviour online only as long as users understand them** and are willing to engage with them. If resources became mandatory for certain audiences to access a platform, some worried that children may let the resource videos play without paying attention.

Content

For adults, the relevance of resources was felt to be dependent on content**. Resources on scams were found to be more useful, especially for those who were less comfortable with technology**. Resources on community guidelines were met by mixed reactions. Some felt these were boring 'terms and conditions', others thought they would be useful to new users and teenagers, as long as they were consulted.

Improvement

As resources were often equated to terms and conditions, **participants felt that clear and accessible writing should be an area of improvement.** Adults suggested that entertaining videos would be needed for teenagers, and this was confirmed by teenagers themselves, who added that resource videos should be short (30 to 60 seconds) and fast-paced.

"[Resources] enable the platform to enforce the rules when someone breaches the standards."
18+, Female, Online diary

"Resources are useful to people who might not have much experience on a website."
66, Male, Adults focus group

"The scam warnings provided by banks before sending money are very helpful to people that wouldn't consider scams."
29, Female, Adults focus group

"Not all people are aware of the exposure that we have [on social] networks and resources are of great help when it comes to informing."
13-15 Male, Children focus group

Parents observed that video-based interventions would work well with teenagers. However, teenagers warned that resources should be short and snappy.

Parents

Parents were split on resources; some felt that videos were an appropriate and engaging format for teenagers, likely to grab their attention. But others questioned whether teenagers would actually consult them.

For some parents, **online behaviour is something that should be learnt, just like one learns to play a sport**. This means users (both teenagers and adults) should learn the 'rules of the game' and platforms should coach and educate them on how to behave.

Parents felt that, **as they stand, resources only allowed platforms to 'tick a box' rather than educating users** about online safety and appropriate behaviour, especially teenagers.

"For me I hate video as an 'intervention'. I much prefer text so I can read rather than watch."
43, Male, Parents Focus Group

Teenagers

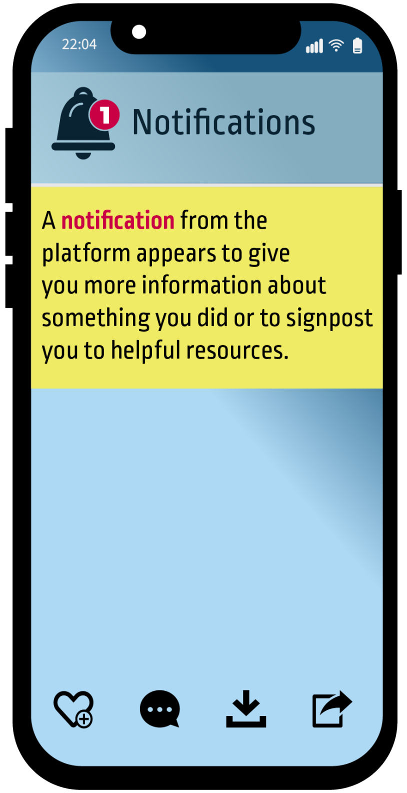
13-15-year-olds liked resources because they felt accurate, explanatory and entertaining. For them**, it was the video format that made them engaging, fast-moving, colourful, easier to process and 'cooler' than the other text-based interventions.**

13-15 year olds saw value in resources because **this intervention clearly spells out what should and should not be done online,** while also giving tips and information on good practices.

Based on these considerations, they described them as 'useful' and 'helpful'. However both older and younger teenagers **felt that resources could easily become irritating if their message was either too childish or long and boring**. They also preferred resources that did not feel too scripted and 'fake'.

"I would probably ignore the "We've updated our community guidelines" intervention as I would expect only minor changes. I would probably open the "How to avoid scammers" intervention the first time to see what information it has to show out of interest, but I would ignore it after that as I would now know the information."
13-17, Male, Online diary

4. Notifications
Definition and examples shown during qualitative stages

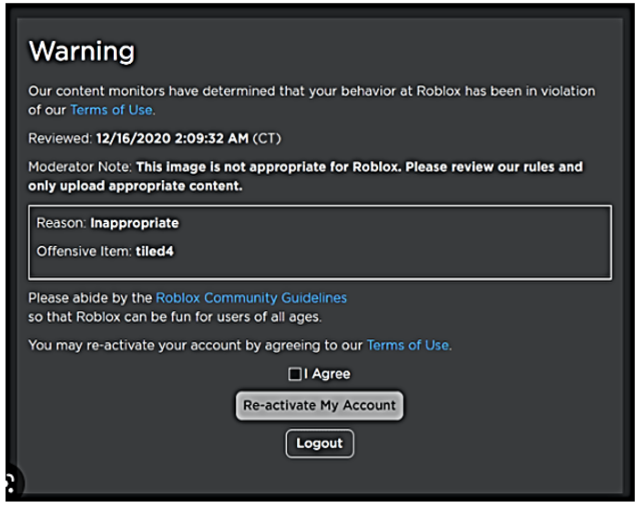


A **notification** from the platform appears to give you more information about something you did or to signpost you to helpful resources.

What is a **notification**?

Notifications

A **notification** from the platform comes in your usual notification feed and gives you more information about something you did (e.g., posted something that broke the platform's Terms & Conditions) or guides you to relevant material.


Warning

Our content monitors have determined that your behavior at Roblox has been in violation of our Terms of Use.

Reviewed: **12/16/2020 2:09:32 AM** (CT)

Moderator Note: **This image is not appropriate for Roblox. Please review our rules and only upload appropriate content.**

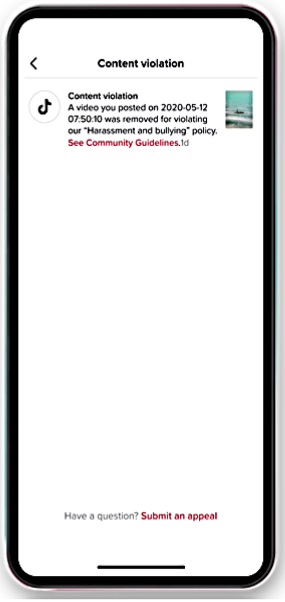
Reason: **Inappropriate**

Offensive Item: **tiled4**

Please abide by the Roblox Community Guidelines so that Roblox can be fun for users of all ages.

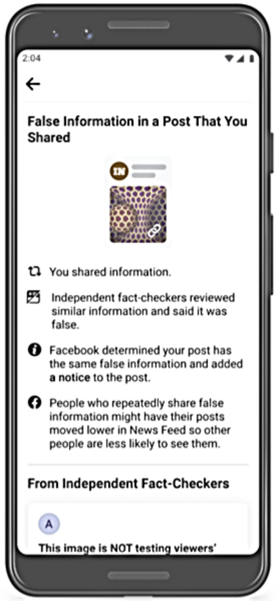
You may re-activate your account by agreeing to our Terms of Use.

□ I Agree


Content violation

Content violation
A video you posted on 2020-05-12 07:50:10 was removed for violating our "Harassment and bullying" policy. **See Community Guidelines.**1d

Have a question? **Submit an appeal**



False Information in a Post That You Shared

You shared information.

Independent fact-checkers reviewed similar information and said it was false.

Facebook determined your post has the same false information and added **a notice** to the post.

People who repeatedly share false information might have their posts moved lower in News Feed so other people are less likely to see them.

From Independent Fact-Checkers

This image is NOT testing viewers'

See appendix for acknowledgements

Notifications

Value

Most adults felt that notifications were not effective, as **they would not stand out among the list of other personal notifications**. By not interrupting online behaviour and needing to be consulted on a separate feed, **they were likely to be missed.** Parents added that children and teenagers would not necessarily understand what they had done wrong and why, without clear and upfront explanations.

Behaviour

Adults felt that for notifications to be successful in changing behaviour and educating users, **they should be read and acknowledged**. This would ensure users understand what they did wrong. Parents suggested that understanding violations in the first place was pivotal for teenagers.

Content

As with other interventions, adults and older teenagers **complained that content can be incorrectly flagged as false information and that the rules around what is flagged and why are blurred.** Parents focused more on harmful and inappropriate content and found notifications to be useful on those cases, although not necessarily effective enough.

Improvement

Participants wanted to see more information about rule violations, alongside reference to the part of the guidelines that was breached. **Some adults suggested an 'appeal' button would help users challenge the platform when content is incorrectly flagged.** Others felt that explanations given via notifications should use simple language to ensure users understand the feedback and take it on board.

"If you're forced to read and acknowledge [notifications] before proceeding then they could help educate the user as to why what they posted was inappropriate."
40, Female, Adults focus group

"These [notifications] are necessary to remove inappropriate material. The first Roblox example takes the additional step to reactivate your account."
44, Male, Parents focus group

"This intervention seems quite useful as it tells you if you have done anything wrong."
13-17, Male, Online diary

"The [notifications] about 'false information in a post that you shared' are usually very annoying because personally I wouldn't share incorrect information, and it often blocks my posts because they're 'false' even though they're not."
16-17 Female, Children focus group

Parents were concerned about notifications not being strongly-worded enough, whilst teenagers found them inaccurate at flagging content.

Parents

For all parents, notifications were not effective as they would not stop inappropriate behaviour. They shared their concerns about their children receiving harassing content online, feeling that social media platforms are 'designed' for adults and constitute a minefield for children.

Parents **thought that notifications could be successful only if strongly worded and if acting as a first warning that would then lead to more serious consequences** (e.g., account deactivation).

They suggested that parents' emails should be copied in any communication acknowledging the warnings. They also thought that rather than referring to the T&C, notifications should mention upfront what was done wrong and why.

"I agree with the principle [of notifications], but when you have kids, AI will specifically draw them in and keep you on the site I think this sadly doesn't work."
46, Male, Parents Focus Group

Teenagers

16-17-year-olds thought **that notifications were useful when informing them about something they did or signposting** them to resources such as for mental health. However, they felt the intervention was often wrongly flagging actions and content and was therefore, inaccurate.

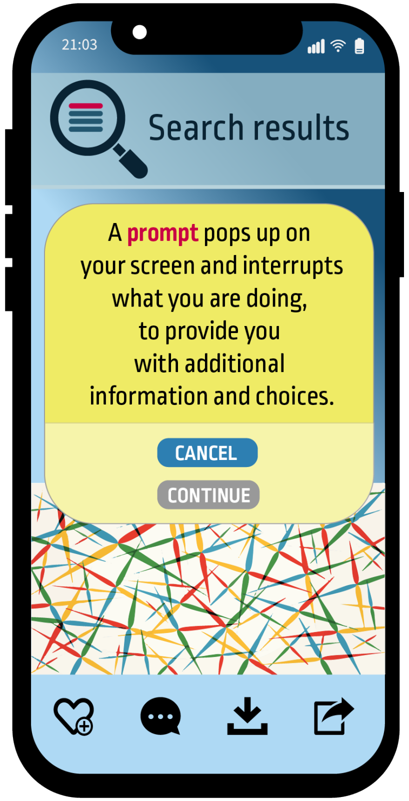
They argued that platforms should be more transparent on how they tag content. They expected notifications to explain specifically why content was being removed, in a clear and understandable way, with links to facts and sources.

They preferred this information to be **upfront, not longer than a couple of sentences** or a paragraph, and with an option to consult more information.

"It's useful in that it tells you what you did. But also, it sometimes malfunctions and tells you, you posted something that breaks the guidelines but is actually fine."
16-17, Male, Children Focus Group

5. Prompts

Definition and examples shown during qualitative stages

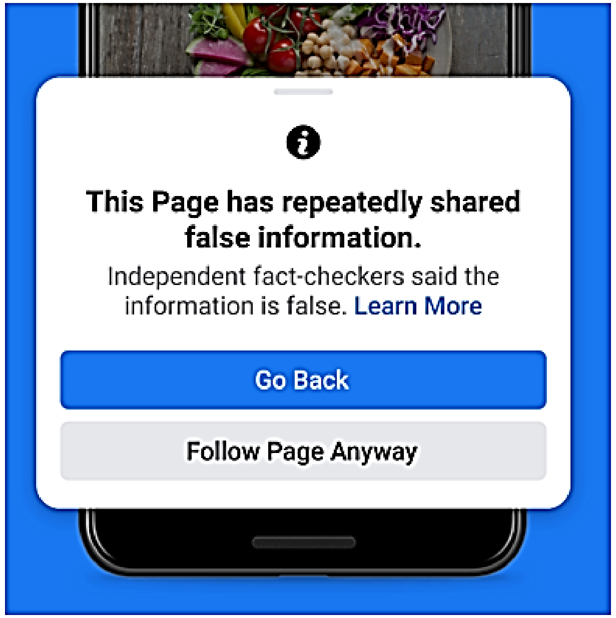


A prompt pops up on your screen and interrupts what you are doing, to provide you with additional information and choices.

Prompts

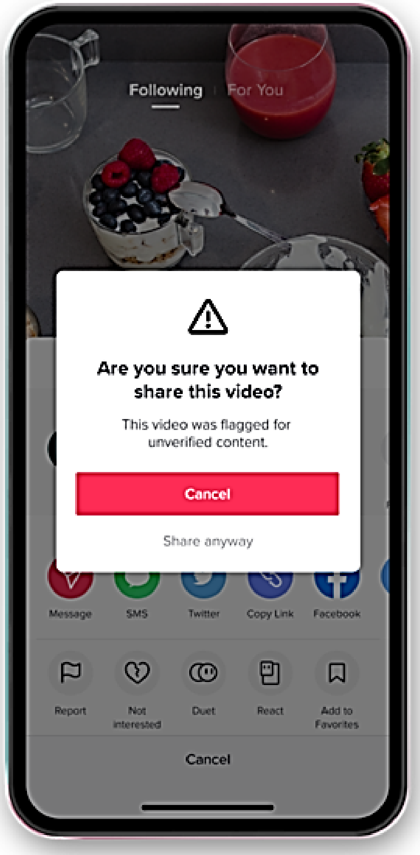
What is a **prompt**?

A **prompt** pops up on your screen and interrupts what you are doing to provide you with additional information and choices. It may also guide you to more information or support.



This Page has repeatedly shared false information.

Independent fact-checkers said the information is false. Learn More



Are you sure you want to share this video?

This video was flagged for unverified content.

See appendix for acknowledgements

Prompts

Value

Adult participants thought that **prompts were informative and enabled decision-making**. Some, however, **argued they wanted to 'decide for themselves',** as they did not trust fact-checkers and did not understand on what principles content is being flagged**. All agreed that sources played a crucial role,** as prompts from banking apps were deemed the most useful and reliable.

Behaviour

Adults had mixed feelings on the ability of prompts to change or educate about online behaviour. **For many, prompts allowed users to 'double check' what they were posting** and potentially do some research. But at the same time, some argued they could be easily dismissed and turn into an irritation that users would repeatedly ignore.

Content

Prompts about false information were praised by some for preventing misinformation and were seen as useful on social media pages sharing news and 'divisive opinions'. **However, there was no consensus on this, as others felt that there was no transparency around fact-checking**. Parents felt more positively towards prompts as they would provide some guidance over online behaviour.

Improvement

For adults, prompts needed: a) more details upfront, instead of a 'learn more' function; b) specific explanations as to what 'unverified' content means; c) bigger and clearer visuals to grab attention. To suit teenagers, adults thought that prompts needed to be succinct, clearly visualise danger and state the consequences of inappropriate behaviour. This was largely confirmed by the teenagers.

"The design is certainly important, red is usually associated with warning and danger, so would be more likely read."
49, Male, Parents focus group

"[Prompts are] irritating and annoying, more than one of these and I would avoid the site / app in the future.'
40, Male, Adult focus group

"Although these can be annoying, I would find them useful. They help curb the spreading of deliberate or accidental false info."
18+, Female, Online diary

"I feel it has useful intent but could get in the way of your work or content you're trying to look at."
13-17, Female, Online diary

While both parents and teenagers found prompts useful, the latter were more sceptical about the guidance role that platforms should have.

Parents

Parents agreed that **prompts were useful** to ensure their children were not exposed to inappropriate content.

Prompts were felt to be especially **necessary on platforms covering news, politics, and opinions**, as in their view these are the places where misleading information may be shared.

Parents also thought prompts **were good at educating teenagers about appropriate behaviour online**, because their young age required guidance. They felt that adults did not need this.

However, **a few worried that interventions like this could generate a 'rebellious' attitude** and entice teenagers to view the content instead.

"I think they might but if pitched wrongly may make a child (or adult) rebel and proceed out of interest."
59, Male, Parents Focus Group

"I don't think I need to be schooled in appropriate behaviour, but it would be useful for kids."
52, Male, Parent Focus Group

Teenagers

13–15-year-olds thought that **prompts were useful in different scenarios**. On social media like TikTok, they would make them stop and reconsider whether going ahead with posting something, whereas on news sites they guaranteed exposure to correct information.

Feelings amongst the 16–17-year-olds were mixed. Some thought prompts were irritating but useful in protecting them. Most, however, questioned who decided what information was false, arguing that this is 'entirely subjective’.

Many appreciated that prompts were short enough and gave the chance to explore more via the 'learn more' feature. However, most would not click through, as they expected the extra information to be lengthy and uninteresting.

"Prompts can stop people from being able to spread potentially dangerous misinformation to large groups."
15-17, Female, Children Focus Group

5. Illegal & harmful content

Findings from text-groups (Adults and Parents only)

Interventions can be used to help reduce the impact of illegal content, but stronger measures to prevent it occurring is what is really wanted.

- The majority of adult participants felt initially that **illegal content should not appear in the first place** and that those who may have shared, liked or searched for content should be banned.
- However, on further consideration, they thought **interventions could be used as a soft initial warning system to educate the user that their behaviour was illegal** (as it might have been accidental). Platforms should then ban users if they continue with their actions.
- There was **far more support for using interventions to offer help and support to those who had accidently been exposed to illegal content**. Platforms were perceived to be responsible for their users' wellbeing.

"I think they should be shown an intervention first time round because it could have been an accident so give them the benefit of the doubt but if it happens repeatedly then there should be a ban."

40, Female, Adult Focus Group

"Perhaps it could function as a first warning type of thing, but definitely ban repeat actions."

20, Female, Adult Focus Group

Interventions were felt to be of particular value in relation to content promoting violence or personal harm

“Only if it is via recognised support networks and not some random social media channel with a 'self-certified' qualification in child psychology.” ***43, Male, Parent Focus Group***

“Yes, because it could trigger or be upsetting to the viewer, especially if they know or have suffered.” ***45, Male, Parent Focus Group***

- **Adult participants were cognisant of the impact of harmful content online**. This was especially related to violent content or information that could promote ‘clear’ harmful perceptions and behaviours (i.e., eating disorders and self-harm).
- The **majority were in favour of the use of interventions to stop instant viewing** of the harmful content and to offer support on the subject matters.
- In the intervention, **participants wanted support resources to appear alongside a warning explaining why the content was harmful.**
- Most felt that, although the platforms held responsibility in relation to harmful content, **the support resources should link to organisations that are experts in the space and hold authority.** Charities were mainly cited here in relation to eating disorders and self-harm.
- If **the content was designed to offer support,** rather than promoting self-harm or eating disorders, **many felt this could be triggering to those who had suffered and an intervention warning them would still be of use**. In this case participants felt that the intervention should clearly outline that this is supportive but may still upset the viewer.

The use of interventions for harmful and illegal information was seen as paramount for child users

- **Participants placed the onus on parents and educators to inform children and teenagers about illegal/harmful content online**. The platforms still held a large amount of the responsibility and should do everything in their power to protect and educate.
- There were **some fears that interventions could tempt younger users to view the content,** but the overwhelming sentiment was that they should be in place to offer some form of intervention, especially to help children who may not confide in their parents.
- **Parents and adults felt interventions should be used here but it would need to be the ones that interrupted the user experience** to block the view of the content (i.e., overlays).
- Second to blocking the view of the content is to provide information on why it has been blocked. The information could educate the user, while relaying the seriousness of the content as harmful or illegal.

"Parents should be informing their children about illegal activities and not abdicating their responsibility to third parties." ***53, Male, Parent Focus Group***

"The children are their users, they should do everything they can to protect them." ***40, Female, Adult Focus Group***

"If my child was experiencing this and didn't [sic] able to confide in myself and wanted to seek support, I'd be happier if an intervention was in place like an overlay that at least gave it verified helplines." ***51,Female, Parent Focus Group***

"[Overlays] Should be used because it stops you temporarily." ***52, Female, Parent Focus Group***

Due to the subjectivity of what could be deemed harmful, some questioned the need for interventions

- **Personal views, misinformation and controversial views (i.e., around news, politics, science and culture) were more contentious for adult participants.** They were unsure of the level of policing needed as they felt it was more subjective and less straightforward than the promotion of violence and self-harm.
- **Adult participants struggled to come to a consensus on the level of harm a personal opinion/experience can cause**. This raised the debate about the autonomy of the user to be able to view information without being influenced.
- Further concerns were raised by a few participants around **who has the power to deem this type of content as harmful.**

"People need to have the choice on whether they want to consume content related to that." ***35, Male, Adult Focus Groups***

"Speeding is illegal. What about legal but harmful? Who decides?" ***72, Male, Adult Focus Groups***

6. Conclusions

Conclusions

- Notifications were perceived as the least valuable and impactful intervention by participants. How they work overlaps with advertisements and platform notifications as well as personal communications, causing them to be largely ignored.
- The other interventions each had strengths relating to specific scenarios online and to different users:

Overlays

- Overlays were seen to be the **most effective** intervention at creating engagement as they stop the user.
- Most felt overlays should be **used where the potential harm is the highest**, this being violent, triggering content and content promoting self-harm. This was considered especially important for the protection of children and young people.
- As overlays can be intrusive, participants felt they must **only be used in scenarios where there is clear risk of harm to the user**, not where there is still debate (i.e., misinformation), as it could be seen as the platform affecting the autonomy of the user to decide what to access.

Resources

- Resources were one of the few interventions regarded as **useful for influencing behaviour** particularly for less digitally literate users and vulnerable audiences (i.e., how to avoid scams).
- **Resources would not change how adults behaved** in terms of inappropriate behaviour, but would be **useful for teenagers**.
- The intervention would need to be re-framed to educate teenagers of the rules on the website, in order to **encourage compliance.**

Prompts & Labels

- Prompts and labels were perceived to **cause minimal interruption** to the user.
- The low levels of disruption makes them **useful in tackling scenarios where the potential harm is ambiguous** and where the use of an overlay may be too much. (e.g., misinformation, paid for posts or controversial opinions where there is debate around the harm to the user.)
- It shows the **platform is doing its duty informing the user** but not overly trying to change their opinion.

Recommendations from the participants of the study

- To improve the visibility of interventions, platforms should look to use brighter and bolder colours. These can be colours that reflect those of the brand but must stand out against the background.
- The design and aesthetics need to be heavily differentiated against other agents using the same mechanisms on the platform (i.e., advertisers) so that it is clear it is an intervention.
- It was suggested that interventions could be colour coded to signal the severity of the content about to be viewed or what the user's behaviour has violated. However, users would need to be educated as to what the different colours mean.
- Interventions need to be used sparingly, an increase in frequency could result in a drop in engagement with the intervention.
- The tone of interventions trying to alert a user to their behaviour should also be reflective of how serious the violation was, while considering the user may have done it by mistake. The platform could take an escalation approach sending a sterner warning if the user repeats their actions.
- The content of the interventions should be neutral and offer support links should the user need it, even in cases where the content is meant to help (i.e., self-harm).

7. Appendix

Acknowledgements

Slide number	Sources
16	1. CBS News Twitter (twitter.com)* 2. YouTube (youtube.com) 3. BuzzFeed YouTube channel (youtube.com) 4. ITV News Facebook page (facebook.com) 5. YouTube (youtube.com) 6. Instagram (Instagram.com) 7. Crime Weekly YouTube channel (youtube.com)
18	1. nike.com 2. o2.co.uk 3. pickmypostcode.com 4. tiktok.com
20	Helping to Protect the 2020 US Elections | Meta (fb.com) Combatting Misinformation on Instagram | Meta (fb.com)
23	https://www.socialmediatoday.com/news/instagram-adds-new-features-to-offer-more-protection-and-well-being-prompts/611112/; https://about.fb.com/news/2021/12/new-teen-safety-tools-on-instagram/ Helpful Search tools for evaluating information online (blog.google) New prompts to help people consider before they share | TikTok Newsroom
26	TikTok Newsroom / TikTok
29	https://en.help.roblox.com/hc/en-us/articles/360020870412-Understanding-Moderation-Messages Adding clarity to content removals | TikTok Newsroom Taking Action Against People Who Repeatedly Share Misinformation | Meta (fb.com)
32	Taking Action Against People Who Repeatedly Share Misinformation | Meta (fb.com) New prompts to help people consider before they share | TikTok Newsroom

*All content was accessed between 14-27 February 2023. On 23rd July 2023, Twitter rebranded to X.

Profile Overview of online text-based focus groups

GROUP	AGE	GENDER	LOCATION
1 Adults 18+	21-74	5 Female / 3 Male	East of England / East Midlands / London / SE / SW / Wales
2 Adults 18+	21-74	2 Female / 8 Male	East of England / London / Scotland / SE / SW / West Midlands
3 Adults 18+	26-72	2 Female / 6 Male	East of England / London / SE / Yorkshire and the Humber
4 Adults 18+	20-71	7 Female / 4 Male	East of England / East Midlands / NW / SW / Yorkshire and the Humber / West Midlands
5 Parents with children aged 13-17	34-59	4 Female / 6 Male	East of England / East Midlands / London / NW / Scotland / SE / Yorkshire and the Humber
6 Parents with children aged 13-17	29-57	3 Female / 7 Male	East of England / East Midlands / NE / Scotland / SE / SW / Yorkshire and the Humber
7 Younger teenagers	13-15	6 Female / 1 Male	East Midlands / London / NE / NW / SE / West Midlands
8 Older teenagers	16-17	3 Female / 2 Male	London / Northern Ireland / Scotland / SE / SW / West Midlands

List of stimulus shown to text-based focus group respondents

	Prompts	Overlays	Labels	Notifications	Resources
Adults group 1	✓		✓		✓
Adults group 2	✓	✓		✓	
Adults group 3		✓	✓		✓
Adults group 4		✓		✓	✓
Parents group 1	✓	✓	✓		
Parents group 2			✓	✓	✓
Children 13-15	✓		✓		✓
Children 16-17	✓	✓		✓	

Profile Overview of online diary respondents – Adults 18+

NO. OF PARTICIPANTS (MALE & FEMALE)	AGE RANGE	INTERNET USAGE
15 (7 Male and 8 Female)	18 - 34	14/15 participants accessed online platforms everyday
15 (9 Male and 6 Female)	35 - 54	14/15 participants accessed online platforms everyday
7 (2 Male and 5 Female)	55+	All accessed online platforms everyday

Profile Overview of online diary respondents – Children 13 - 17

NO. OF PARTICIPANTS (MALE & FEMALE)	AGE	INTERNET USAGE
2 (1 Male and 1 Female)	13	All accessed online platforms everyday
2 (2 Female)	14	
3 (3 Female)	15	
4 (1 Male and 3 Female)	16	
1 (1 Male)	17	