Avatar Methodology

Pilot study

A report for Ofcom

REVEALING REALITY

Ofcom

About Revealing Reality

Revealing Reality is an independent social research agency, working with regulators, Government and charities to provide independent and rigorous insight into young people's online behaviours and experiences.

Studying how the digital world is shaping people's lives is something we do every day. We have been tracking children's media use and the impact it has on them for the past nine years as part of Ofcom's Children's Media Lives research, and we've recently conducted work for Ofcom about the harms children experience online. In the digital realm, we have also done detailed qualitative behavioural research and explored the effect digital design has on people's experiences.

Visit www.revealingreality.co.uk to find out more about our work or to get in touch.

Background and context

Qualitative research has shown that children under 13 in the UK are being exposed to potentially harmful content.[1] Ofcom identified a gap in the available data around understanding the extent of exposure of under 13s to this potentially harmful content. Under the requirements of the current draft of the Online Safety Bill, services that are likely to be accessed by children will be required to consider the risks to children in different age groups from harmful content, and to use systems and processes designed to protect children from encountering harmful content. Understanding more about what children are exposed to online will aid Ofcom in delivering its current media literacy duties.

Ofcom commissioned Revealing Reality to conduct a pilot to test the feasibility of understanding harmful content to children aged under 13 using 'avatars': accounts set up on several online platforms by researchers, modelled on the behaviours and interests of real children. Using the avatar accounts, the researchers were able to record the content they are exposed to on different online platforms.

The purpose of this research pilot was to test this method for understanding the online content that children under the age of 13 are likely to be exposed to, in order to evaluate whether this methodology could increase understanding of situations that children could find themselves in online - especially exposure to potentially harmful content.

This report outlines the key learnings from this pilot study and considerations which could help Ofcom consider use of avatars in the future.

Methodology

Ofcom commissioned Revealing Reality to run an extensive scoping exercise in which the legal, ethical, and practical considerations of running avatar research were considered. Following this review and the subsequent decision by Ofcom to run pilot avatar research, Revealing Reality set up 15 online accounts based on five real children interviewed by researchers. These interviews included the submission of five minutes of screen recording, and screenshots of some key settings across the different platforms the children used. This included screenshots of privacy settings and account set up information such as date of birth.

Using this data, researchers recreated (within the parameters set – see 'Rules of Engagement' pull out box below and Annex 2) some of the children's online habits and interests on 15 different accounts across nine different online platforms used by the children. The research was conducted on platforms that Ofcom research shows are used by children aged 12 and under.[2] These platforms were Twitter, Twitch, TikTok, YouTube, YouTube Kids, Roblox, Facebook, Instagram, and Snapchat.

Prior to the avatar tracking phase in the research, Ofcom informed the platforms involved about the research and published a transparency notice on the Ofcom website to inform the public about the research.

Coding content

Researchers spent between 10 and 15 minutes a day per account, following a pre-agreed set schedule unique to each avatar. These schedules reflected the interests of the real children, as gathered in the qualitative phase. All avatar tracking sessions were screen recorded to allow for analysis and coding.

Once data was recorded, it was coded using a red, amber, yellow, green (RAYG) framework to indicate different levels of potential harm to children (see Annex 3). The content was coded as follows:

- Red content was very likely to be harmful to children and adults.
- Amber content was likely to be harmful to children.
- Yellow content was age-inappropriate content which may be harmful for some younger children.

[1] https://www.ofcom.org.uk/research-and-data/media-literacy-research/childrens

[2] https://www.ofcom.org.uk/research-and-data/media-literacy-research/childrens

    - Whether a piece of content was age-inappropriate was assessed based on the actual age of the child using content parameters guidance from organisations (such as the BBFC).
- All other content was coded as green, which was unlikely to be harmful to children.

This system was used to minimise subjectivity between the researchers when coding and provide an accurate standard of analysis. However, due to the qualitative nature of the research, it wasn't possible for the coding system to be completely objective. Revealing Reality put measures in place, such as a coding guide and weekly analysis sessions, to cross-check coding decisions and ensure that the deployment of the method remained systematic.

Safeguarding of researchers

Throughout the project, the wellbeing of the researchers running the avatar accounts ('avatar managers') was taken very seriously, with several safeguarding measures in place.

Avatar managers were briefed before the project to explain the types of hazards and content they may see during the project. They were then asked to give informed consent to take part in this project and were informed they could be moved onto other projects with no consequences. Avatar managers were encouraged to share any concerns with the senior team immediately, especially if they felt this work was affecting them. Daily check-ins had a standing agenda item regarding each avatar manager's wellbeing.

To minimise risks to any other colleagues, content was reviewed in private spaces where avatar phones were not visible to others. During the project, no high risk or potentially illegal content was encountered. However, strict processes and safeguarding protocols were put in place in the event that avatar managers did come across this type of content.

Prior to the research starting, Revealing Reality conducted a risk assessment in case a member of the team came into possession of high risk or potentially illegal content. While the risk severity was high, the likelihood of the risk given the mitigations was low.

Overview

The avatars were based on real children interviewed during the qualitative phase of the research. The avatars were registered on the platforms the children used and registered as the same age as the real child's account, which often differed to their real age. They then followed similar interests and accounts. The research team recognise that the creation of an account by a user below the minimum age is a breach of such platforms' terms of service, and could pose risks to the child concerned, particularly if they are purporting to be over 18. Therefore nothing in this report should be taken as an endorsement of creation of such accounts.

Pseudonym	Real age	Platforms	Registered age/s
Sophie	4	1 platform	4
Danny	7	2 platforms	18+/ 3 with 13+ experiences enabled
Oscar	10	4 platforms	18+
Poppy	12	5 platforms	18+
Ali	12	3 platforms	18+/14/18+

Key learnings

This was a small-scale pilot study, designed to test the avatar methodology rather than deliver robust findings on the content children see online. For this reason, this report primarily focuses on findings relating to the feasibility of the methodology rather than content encountered by the avatar accounts.

The pilot explored what children under 13 could or are likely to see based on some of their online interests and behaviours.

Avatars are a valuable tool to understand experiences across the range of online platforms children use

Avatars have been used as part of journalistic investigations[3][4][5] and by charities[6] to give an insight into potential risks children could come across online. However, this is the first time Ofcom has tested whether this methodology could help to understand what a child is likely to see based on their current behaviours and interests.

This particular methodology has been developed by Revealing Reality in collaboration with Ofcom to ensure that the avatars were run systematically, ethically, and as 'true to life' as possible within those parameters. While the body of this report primarily focuses on key learnings from the pilot and considerations for the future, it is important to note that avatars could provide valuable insight into specific platforms that other research methodologies (such as interviews and passive tracking) may not be able to provide. The pilot has demonstrated that the avatar methodology is a useful research tool that can help increase Ofcom's awareness of the different online spaces children may be part of, as through this method the avatars were able to collect data about the type of content children were exposed to online.

Through avatars, Ofcom could gain a deeper understanding of what real children may see online and what content they could be exposed to. There is also the potential for avatars to provide 'real time' context such as the pathway to how a piece of content is encountered by an avatar, or how frequently different types of content are encountered by an avatar. If used on an ongoing basis, avatars could also provide Ofcom with a greater understanding of online trends as they play out on different platforms.

Avatars could be used by Ofcom to understand different platforms' changing online safety measures, e.g., age assurance measures or content warnings.

Whether avatars are a robust method for understanding what children are exposed to online is dependent on how closely the avatars can mimic real children's behaviours. How well this can be done is dependent on a few factors, including:

- The quality and quantity of data obtained from the qualitative research among children to inform the avatar tracking stage.
- Ethical and procedural limitations on what the avatars can do on each platform.

[3] https://www.businessinsider.com/researcher-claims-her-avatar-was-raped-on-metas-metaverse-platform-2022-5?r=US&IR=T

[4] https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/6C0C2040-FF25-4827-8528-2BD6612E3796

[5] https://www.bbc.co.uk/news/uk-61813959

[6] https://5rightsfoundation.com/in-action/new-research-shows-children-directly-targeted-with-graphic-content-within-as-little-as-24-hours-of-creating-an-online-social-media-account.html

The validity of an avatar would be improved by collecting more data to inform its ongoing behaviour

The avatars were limited by the amount of behavioural data that can be captured about a real child's activity, due to ethical and logistical considerations. The pilot avatars were constrained to behaviours that were described during interviews or exhibited in a screen recording of 5-10 minutes. These behaviours were then scaled to create the four-week behaviour schedules for avatar managers to replicate. The behavioural schedules were fixed to the behaviours the children exhibited during the qualitative phase. As avatar managers did not deviate far from these behaviours, they did not always know whether new recommended content would be of interest and, in line with the protocols of the pilot study, assumed that nothing except the known interests of the real child would be of interest. The tracking ran for a four-week period to allow avatars to 'acclimatise' to each platform. During this period, avatar managers observed the content served to the avatars based on their initial account inputs and saw content change as avatars became more established on each platform. It appeared to researchers that some algorithms were responding to the avatar's input, as the type of content served changed over time, demonstrating the value of accurately re-creating unique real behaviours.

The qualitative data obtained for this pilot provided an insight into how children use their different platforms. However, to robustly measure the online experiences of children under 13, the data set of online behaviours would need to be larger. Any future study could consider gathering longitudinal data from children to inform the ongoing behaviour of the avatar. This might include screen-recording children's behaviour for longer. However, increasing the data set raises important legal and ethical considerations. Longitudinal data collection, such as using screen record over several days, could be deemed to be intrusive to the participant and other users.

An alternative approach to collecting data for the avatars, which would be less intrusive, could be to collect shorter screen record clips as part of an ongoing task list, on a more regular basis. These would supplement the initial interviews and contribute to a larger data set that would enable a more accurate avatar methodology.

There is a need to comply with legal requirements, including in relation to data protection. Designing avatars requires collecting data (including personal data), which is detailed enough to inform avatar schedules, whilst minimising the potential intrusion caused to research participants and third party platform users by this process.

On balance, the small-scale data set used in this pilot would not be sufficient to run avatars in a robust manner beyond a pilot. Taking into consideration the limitations on using more screen-recording data, any future research using an avatar methodology should ensure that the avatars are based on more information from the child research participant. For example, by using a diary task and/or collecting data over a longer period.

The necessary protocols and Rules of Engagement meant the ability to replicate children's behaviour varied by platform

The Rules of Engagement (Annex 2) were developed in line with the ethical considerations in order to minimise the potential impact of the avatar's actions on other users online and minimise the risk of interacting with children and private information. There were some limitations on accurately recreating the children's behaviours due to restrictions around what content the avatars could engage with. Avatar behaviour was limited by the rules relating to 1000 followers, 1000 'likes', and not following private accounts that seemed to belong to children. These rules meant that some online spaces were inaccessible to the avatars, which substantially limited the ability to recreate certain aspects of children's online behaviour on certain platforms.

The impact of the '1000 rule' varied across platforms. Avatars on platforms where a lot of content was produced by larger creators tended to be more reflective of children's online behaviour, as children on these platforms typically only engaged with larger profiles and content creators. Therefore, avatar managers were able to engage with the majority of the accounts and more accurately recreate their behaviours.

Furthermore, some platforms have features that allow users to hide 'like' counts under their posts. In these cases, avatar managers took a risk-averse approach and did not engage with the content. However, there is potential that more platforms could adopt this feature and according to the '1000 rule' would restrict the behaviours of avatar managers further.

Platforms that emphasised peer-to-peer exchange, or where children more regularly engaged with their peers to find new content via share functions, were more limited by the Rules of Engagement. This was true for the '1000 rule' limiting engagement with content and accounts with under 1000 likes or followers, and the risk-averse approach of not engaging with users who appeared to be under 25. Where children were using accounts to engage with peers, researchers were unable to engage with these accounts and were not able to replicate this key part of children's online behaviour on these platforms. Furthermore, some behaviours were not appropriate to replicate due to logistical rather than ethical reasons. For example, many of the children the avatars were based on spent several hours online each day. For the purposes of the pilot, it was not feasible for the avatar managers to replicate how long the children spent online.

Avatars can be an ethical methodology to understand what children may be exposed to

With strong protocols in place, avatars can be a useful, ethical method for providing insight into what children may be exposed to online based on their interests and behaviours.

The avatars' hobbies and interests seemed to play a key role in shaping the content they were exposed to. In some instances, algorithms were providing content that appeared relevant to the avatars' interests, but which were in fact potentially hazardous content. For example, one avatar saw a short-form video appearing to relate to arts and crafts, which in fact showed creative ways to roll cannabis cigarettes. Another avatar saw a short-form video which appeared to relate to cooking and baking but featured cookies with explicit and sexually suggestive language written in frosting.

Avatar managers observed that avatars were exposed to potentially harmful content more frequently when using a short-form video format than text or post-based functionalities, as well as the differences observed around different interests outlined above.

Using the avatar methodology, it is possible to draw insights from within a platform (where an avatar can ethically replicate the behaviour of a child) of the factors that could place a child more at risk. By running several avatars on the same platform, with a large enough sample, it would be possible to draw insights into the impact of an avatar's behaviour on the type of content being seen. Similarly, changing and comparing variables such as the age (both the age set up on the account, which was often older than the child's real age, and the actual age of the child the account is based on), gender, expressed interests, behaviours and settings or parental controls within a platform could all help understand how to reduce risk. In the pilot, all but one of the children interviewed did not use parental control functions, which was then replicated in the settings of the avatars' accounts.

Avatars are not suitable for robustly measuring differences between exposure to content on different platforms, because children's behaviour varies by platform

Despite being able to measure exposure to potentially harmful content within a single platform, this pilot has demonstrated that an avatar methodology would not be a robust method to compare the content children are exposed to across different platforms.

The ability for avatars to accurately replicate children's behaviour varies substantially across each platform and each platform type, due to the different way each platform is used (i.e., the different functions available) and the ethical limitations on what behaviours can be replicated. As explained previously, avatars were not able to engage with peer-to-peer content due to the '1000 rule' in the agreed Rules of Engagement. Therefore, the avatars could not replicate behaviours on platform functionalities that focussed more on peer-to-peer contact. As such, the record of the avatar's exposure to potentially harmful content is limited to content via other functionalities and therefore likely to be less accurate for these platforms. In contrast, on a platform where children were primarily engaging with content mainly from larger content creators, the avatars were more likely to replicate the child's real behaviour. In summary, the protocols necessarily limited avatar behaviours which influenced their experiences on some platforms more than others. This would suggest that this avatar methodology is not the most suitable to draw cross-platform comparison.

Additionally, the current coding process involves coding individual pieces of content according to the perceived level of hazard, (see page 3). This means that it is not clear how best to compare exposure on different platforms where the number of pieces of content consumed in a period could be very different. For example, a child may only watch one or two long form videos in 15 minutes, but may read a high number of text-based comments in the same period. Solutions could be developed such as coding content every 30 seconds throughout the period regardless of the source (so that a longform video receives multiple codes, while many text-based comments would be skipped). However, this would be difficult to implement, and a single video might still distort the data.

Due to the number of unknowns about what 'inputs' shape a user's experience on a given platform (i.e. how the platform algorithms function) and differences across platforms, avatars are unlikely to provide a suitable cross-platform measure for prevalence of potentially harmful content, but instead allows insight into what children *could* be seeing online on particular platforms based on their unique behaviours and interests.

There was value in having a coding category for 'age-inappropriate' content to add greater nuance

Coding the content was not straightforward, and it was difficult to capture nuance and limit subjectivity around coding decisions. For example, a joke playing on racial or gendered stereotypes is subjectively harmful depending on the child's existing knowledge of the world. Having a four-point scale for coding, ranging from red to green, made the task of coding more straightforward, as it enabled the team to capture more nuanced data on what the avatars were seeing. However, it was still difficult to consistently classify content into these categories in the absence of more detailed guidance.

The margin between what might be classed as age inappropriate ('Yellow' content) and potentially hazardous ('Amber' content) was not clear cut and required more analysis between the avatar managers. Hazards sit within a spectrum and the 'cut off' point for what may be classed as potentially harmful is dependent on context. For example, to be able to code content, there was a requirement for researchers to hold some understanding of internet slang or references that were not always obvious to some. Some age-inappropriate content had subtle coded references through emojis, such as the green falling leaf emoji being a reference to cannabis (this was coded as age-inappropriate as the reference was very subtle and the context around the video did not reference any other similar risky behaviours). Without this understanding, there is a risk that content could be mis-coded.

Changes could be made to the design of the coding framework to account for contextual information

The coding framework, as designed, did not account for contextual information such as how the avatar came to be served the content being coded, e.g., whether pushed via a notification or served in a feed. Avatar managers observed that this information felt pertinent to how it might, if served to a real child, impact them, and therefore might merit being captured by the coding framework in future work. For example, one of the avatars had seen a video including 'rude snacks' which featured sexually explicit text written on the cookies – this content was encountered after avatar managers searched for cooking related videos. The current method and coding framework does not account for instances like the example above, where the avatars came across hazardous content somewhat unexpectedly. The research team know from other research that when harms are unexpected the impact on the child is likely to be greater.[7]

The coding framework also did not readily identify trends or commonalities across types of content, especially where they might be below the threshold of being coded as potentially harmful (i.e. coded as green). For example, where avatars might be served a high concentration of content on a theme – such as misinformation about the benefits of going to school – which, while coded as green, may have a cumulative harmful effect on children. The fact that platforms were recommending these clustered or high concentrations of content types

[7] https://www.ofcom.org.uk/research-and-data/online-research/keeping-children-safe-online

again felt pertinent to the coders as a relevant feature but was not easily captured by the coding framework as currently designed. Also the research team know from other research that the cumulative effect of repeatedly seeing borderline harmful content is likely to have a greater impact on children than only seeing this type of content once or twice.[8] The way in the which the content reaches the avatar could also have an impact on the level of potential harm e.g. viewing a video may have a different impact to receiving a direct message.

Additional nuance between photo and video content may be particularly useful because multiple hazards may appear throughout a video at different times. The current coding framework classes videos as a single piece of content, which neglects to consider how videos often feature more than one scene, where different moments within the same video could be coded differently. The approach taken counted each video where a hazard occurred with a single yellow or amber rating (whichever was the highest). Not coding multiple moments within a video could misconstrue the scale of potentially harmful content within it, if for example a single piece of content contained multiple yellow or amber elements. Therefore, it may be worth considering coding multiple frames within a video, or, as suggested above, coding on a more frequent time basis to capture different moment within a video that the child would encounter.

Similarly, any future study may want to consider capturing the frequency of types of content encountered within the 10 – 15-minute tracking period which may enable analysis of a cumulative risk rating for example, repeated exposure to borderline content. Including this within the coding framework could begin to provide a more accurate image of the potential level of harm a child could be exposed to within a platform.

Additionally, the coding framework does not include other forms of media encountered whilst navigating to content, such as thumbnails of content. Considering the amount of data already being gathered, and the nature of coding the data, it was decided that during the pilot this thumbnail content would not be coded. There would be value in considering this additional content, as children would still engage with this whilst using platforms. However, coding further content would take additional resource, which must be considered.

As a general point, there are improvements that could be made to the coding system to add greater nuance and clarity. However, many of these improvements relate to coding a greater quantity of content – which has implications for time and resource.

The avatars were exposed to potentially harmful and age-inappropriate content

Despite not being the focus of this report, findings relating to content encountered offer a glimpse into the potential of a more developed avatar methodology. As outlined above, because this reflected the behaviours of the research participants in the earlier part of this study, most of the avatars were registered with an age of 18+ across the different platforms. The accounts that were registered as under 18 included one account on one platform with an age of four, one account on one platform with an age of three and one account on one platform with an age of 14. Avatar managers observed that all except one avatar (registered as aged four) were exposed to age-inappropriate and potentially harmful content.

- Most of the content seen by avatars was coded as green, followed by some yellow content featuring swearing or content that was considered age-inappropriate for younger children (aged 8 and under).
- There was less amber content, which included violence, sexually suggestive material, suicide ideation, references to mental health issues, and in a few cases, references to the use of illegal drugs such as cannabis.
- The avatars saw no content that would have been coded as red.

As outlined throughout this report, different behaviours and demographic information can be seen to result in different exposure to different types of hazards. The avatar managers observed that all of the older avatars (based on a child aged over 10 in real life) had been exposed to sexually suggestive content, with the male avatars seeing content linked to popular adult content subscription sites. The oldest male avatar saw the

[8] https://www.ofcom.org.uk/research-and-data/online-research/keeping-children-safe-online

highest proportion of sexually suggestive content compared to his counterparts, in line with his real-life behaviours and interests to similar content.

The oldest female avatar saw a greater amount of content relating to mental health conditions. The child this avatar was reflecting had expressed an interest in "venting" on one platform, where users would publicly post about topics they were unhappy about, from homework to mental health. This online interest was confined to this specific platform and was not seen across the child's other accounts. However, despite not searching for, or following, any content related to mental health on other online platforms, the avatar was served similar content across different accounts. This may be related to wider platform algorithms. This avatar encountered content varying from material supposedly encouraging recovery and suicide prevention, but with potentially harmful romanticisation or fixation on death, to self-diagnosis and broader discussions relating to mental health.

Avatar managers also observed one male avatar being served more age-inappropriate and potentially harmful content throughout the course of the tracking period, despite the input behaviours remaining largely similar throughout the weeks. For example, the avatar had an interest in history related content, and towards the end of the tracking period, had been served some content with xenophobic undertones.

The following findings should be considered alongside the rules of engagement imposed on the avatars. Importantly, their inability to interact with users with under 1000 followers, those appearing to be under 25 or private accounts, which limited the content the avatars were able to engage with. Therefore, where these restrictions led to a larger gap between the avatar behaviour and that of the real child the approach is less reliable. In terms of the content observed the avatar was not able to follow particular pathways e.g. content being passed between peers.

Avatar managers observed that platforms with primary functions related to short form video or content discovery pages tended to result in more content being coded yellow or amber than on platforms that primarily revolve around a feed of content from accounts the avatar had chosen to follow.

Avatars operating on platforms with a heavier emphasis on 'chat' and peer-to-peer exchange tended to come across a lower rate of 'yellow' and 'amber' than other forms of content. But, as noted above this may have been impacted by the restrictions placed on the avatars and the limitations of this pilot.

During the tracking phase, there were three instances across two platforms of prompting notifications flagging that a piece of content was age inappropriate. In these instances, no ID was provided, and the avatar's age was neither verified nor disputed by the avatar manager. The different circumstances in which this occurred are outlined below:

- A platform required the avatar to produce ID for age verification to continue accessing certain features and to access specific content after the avatar searched for sexually suggestive content, in line with the real child's behaviour.
- A platform prompted age verification in response to the avatar attempting to view a piece of content that had been suggested via a push notification. Upon opening the notification, the video was flagged as potentially inappropriate for younger audiences and the platform would not allow the avatar to view the content.
- A platform also prompted age verification for certain features of the platform, despite being registered as over 18, after possibly identifying behaviours that the platform thought suggested the avatar's account was likely to belong to a child.

Considerations for future work

Through undertaking this pilot study, the research team have identified five key considerations for future work:

1. Avatar behaviours are heavily dependent on the quantity and quality of the data provided

The avatars were limited by the amount of behavioural data that was captured about a real child's activity, due to important ethical and practical considerations.

While the qualitative data provided an insight into how children use their different platforms, to robustly measure the online experiences of children under 13, the data set of online behaviours would need to be larger. However, there are important considerations and limitations around the level of data that should be collected about a child. Therefore, there is a delicate balance to be struck between collecting enough data to improve the accuracy of the avatars and ensuring the research method is not unduly intrusive as to the privacy rights of individuals including children.

2. Time required to track and code material

The amount of time needed to track and code material is one of the biggest considerations encountered by avatar managers. Avatar managers allocated three hours daily to run the avatars (for 10–15 minutes per account per day), which amounted to 300 sessions. While the managers had accounted for this time, the additional time used to *code* the 8000-10,000 pieces of content surpassed expectations. Time would need to be allocated accordingly when replicating this method.

Short-form video content took longer to code than text-based content, due to the additional context managers need to provide, and the fact that sometimes they needed to be watched more than once if there were additional layers of content (i.e. audio unrelated to the video, captions, or text overlay). In general, text content was often quicker to code than any other form, and longer video content was less time consuming than short form video due to less jumping between topics and lower likelihood of containing several layers of content at once (i.e. split screen, unrelated audio, text overlay).

3. Understanding of the avatar managers

There is substantial benefit in avatar managers having an up-to-date knowledge of current trends in the content liked and shared and language used by children, as this knowledge helps to provide greater understanding of content and its potential for harm. A key example of this being where emojis are used to indicate drug use, or acronyms relating to self-harm or eating disorders are coded to avoid being filtered by a platform. There is a risk that if the full team lack this understanding of youth culture and internet slang, harmful content could be overlooked or miscoded. This also highlights the necessity of peer reviewing classifications as a team and having discussions to assess potentially hidden meanings and potential harm.

Avatar managers would also benefit from understanding and experience of gaming platforms. Avatar managers assumed a more 'active' role on gaming platforms compared to social media platforms, as running an avatar here required managers to play specific games. Understanding how to play these games would allow the avatars to run more accurately to the child's behaviours.

3. Platforms hiding 'like' counts

This was encountered by the avatar managers on a few occasions, who followed the risk averse approach by not engaging with this content to follow the required protocols. If this trend amongst social media platforms continues to grow, it may hinder the ability to follow the Rules of Engagement and the mitigations would need to be altered.

4. The methodology is not appropriate to compare hazard exposure rates across platforms

By running several avatars on the same platform, it was possible to observe different levels of exposure to different types of hazards depending on the avatar's behaviours and interests, which could be scaled up. Although avatars can measure exposure to potentially harmful content within a platform, on balance, this pilot has demonstrated that the avatar methodology would not be a robust method to compare the content that children are exposed to across different platforms. Due to the unique behaviour schedules for each platform and the ethical limitations about what behaviours could be replicated, how accurately avatar managers could replicate children's behaviour varied across each platform type. This level of variability meant that cross-platform analysis was inappropriate.

5. Avatars can provide insight into what hazards children could be exposed to online, but they are not an appropriate tool to determine whether this is harmful

Avatars can provide an insight into the hazards (the potentially harmful content) that children may be exposed to, but they cannot draw conclusions about the impact of these hazards on children or the harm they may encounter as a result. As a method, avatars are limited to the specific platform they are being run on and while they can identify hazardous content and interactions encountered on specific platforms, the method does not account for understanding other variables that may lead to harm, such as factors related to the user (e.g., characteristics and circumstances) and exogenous factors that may impact the user's experience (e.g., societal context).

6. The study and coding framework could be adapted to capture further detail and nuance around types of content including the contexts around them

The coding approach used treats all content the same, and while it offers an effective way to analyse large amounts of data, which is a scalable and consistent, this does compromise insight into some of the nuances of the content encountered. It also did not capture other context around a piece of content – such as the frequency of types of content encountered within the 10–15-minute tracking period and other types of content encountered on the way to the piece of content (such as thumbnails).

Furthermore, due to the unique nature and different types of content encountered on each platform, there is value in having specific coding protocols for each platform. This would consider platform specific context such as content type and length.

These would all be valuable additions for adding greater detail and nuance to the framework. However, this would mean coding a greater quantity of data, which comes with implications for time and resource.

Annex 1: Detailed methodology

WC 16/01/23 - 06/02/23	WC 06/02/23	WC 13/02/23 - 06/03/23
In-depth interviews with children and their parents to understand online behaviours and wider interests, including sharing their screen live.	Each avatar was set up based on a respondent from the qualitative phase.	Tracking data was collected daily for a period of four weeks for three hours.
Interviews explored search history, profile details such as biographies and profile settings including privacy and age restrictions. Qualitative research also included screen recording sessions performed independently and sent to researchers, in which children were told to use their social media platforms as they do normally.	Children's avatars replicated the platforms used and account settings. The profiles of the Avatar accounts were based on the children's profiles, but in such a way that the child could not be identified from this. Children's interests were engaged with by following the same accounts where permissible by the 1000 engagement rule and following some similar accounts when not possible.	10-15 minutes per day was spent tracking each avatar account. Tracking took place between 3pm and 6pm to replicate the after-school use of media by the children interviewed. Using the qualitative data collected, behaviour schedules were created to replicate observed behaviours in a controlled manner to that ensured all tracking protocols were abided by. Screen recordings were stored on hard drives kept locked away when not in use.

Annex 2: Rules of engagement and platform mitigations

Protocols and mitigations

Type of engagement	Protocols and mitigations
Follower limits	Avatars will only follow accounts with a minimum of 1000 followers. • Avatars may only follow 'child-like' accounts if they are public **and** have at least 1000 followers. Avatars may follow 'adult' accounts if they are private or public if they have at least 1000 followers.
Potential engagement with children online	Avatars may only follow 'child-like' accounts if they are public **and** have at least 1000 followers. If there are any doubts as to whether the account holder is a child, avatar managers will take a risk-averse approach and apply a 'challenge 25' principle – if the account holder appears (from the information and content present) under 25, the avatars will cease further engagement. If, after applying this principle, there are still any doubts, avatars will also stop any engagement in line with being risk averse.
'Liking' and commenting posts	Avatars may like posts (if in line with the real child's behaviour) if the post has over 1000 likes. Comments are neutral emojis or statements that are pre-agreed with Ofcom, such as the laughing face emoji.
Replying to direct messages	Avatars may reply to direct messages if the account holder is presenting as an adult (see row on 'potential engagement with children online'). Messages are neutral and pre-agreed with Ofcom, for example, replying, "what is this?". Following this initial reply, avatars will not engage further with the account.
Reporting content	Avatar managers may report legal but age-inappropriate content if this is in line with the real child's online behaviours.

As there is a lack of sufficient information to determine whether a user is an adult or a child on Roblox, there will be no engagement with accounts and profiles on Roblox. Avatars may also only play popular games that have over 1000 live players. While there are multiple servers within each game that have limits of under 1000 players, the avatars would not be engaging with any other players and the risk to others is very minimal.

There is also a lack of sufficient information available on Snapchat to determine whether a user is a child or adult. As information on the number of subscribers/friends is not readily available to other users, avatar managers cannot apply the '1000 rule' when deciding any potential engagement with children online. As such, avatars will not engage with any private accounts on Snapchat and will only 'subscribe' to verified accounts.

Annex 3: Risk coding framework

Category	Description	Examples
Red content – very likely to be harmful	• Content that is illegal or potentially illegal, or extreme content that is very likely to cause harm (for example, because it may promote a potentially illegal act)	• The content amounts to incitement of a child to engage in sexual activity • The content is pornographic • The content might assist, promote, encourage or provide instructions for suicide • The content might promote, encourage or provide instructions for deliberate self-harm • The content includes threats to kill or of violence against an individual or individuals • The content promotes or supports a proscribed terrorist organisation • The content amounts to an offer to sell or supply a firearm or other prohibited or controlled weapon (including where it would be illegal for a child to buy such an item)
Amber content – likely to be harmful	• Content that is likely to cause harm in some contexts (e.g. to certain users or cumulatively over time). • Contact that is likely to cause harm (e.g. to certain users or cumulatively over time).	• The content might promote, encourage or provide instructions for eating disorders or behaviours associated with eating disorders • The content depicts violence against individuals • The content appears likely to incite hatred against people based on protected characteristics • The content is abusive and targets protected characteristics • The content is likely to incite violence against individuals • Information that is likely to be harmful to the child's health (may be red depending on severity)
Yellow content – Age-inappropriate content, may be harmful for some children	• Content that may be uncomfortable or unsuitable for audiences under 8 (e.g., may not be harmful for a 12-year-old but may be inappropriate for an 8-year-old). • Content that is age inappropriate and may be harmful in some contexts but is generally unlikely to cause harm.	• The information appears to promote fake news • Moderate bad language • Sexual content discreetly implied (ie: use of a euphemism)

Green content – unlikely to be harmful / not harmful	• Content / contact that is unlikely to cause harm to most users	• Cute animals • An adult woman swimming on a beach • Non-harmful pranks

Annex 4: Thematic coding framework

This framework was developed by Revealing Reality and Ofcom to support the analysis and coding of yellow, amber and red content by theme. It was prepared taking into account a version of the Online Safety Bill available at the time of research development and other relevant data sources, such as British film classifications.

1. Mental health

Content describing, referencing and depicting mental wellbeing, illness and mental health conditions.

a) The content appears to depict mental health conditions
b) The content appears to encourage self-help and recovery
c) The content might assist, promote, encourage or provide instructions for suicide
d) The content might promote, encourage or provide instructions for deliberate self-harm
e) The content appears to depict deliberate self-harm
f) The content appears to depict suicide
g) The content appears to attempt or encourage mental illness diagnosis

2. Body image

Content describing, referencing, depicting feelings and thoughts around a person's body. Body image may range between positive and negative.

a) The content might promote, encourage or provide instructions for eating disorders or behaviours associated with eating disorders
b) The content might depict eating disorders or behaviours associated with eating disorders
c) The content might promote, encourage or provide instructions for weight loss and food restriction
d) The content might depict weight loss and food restriction
e) The content advocates for body positivity
f) The content might promote, encourage or provide instructions for health, exercise and fitness
g) The content depicts non-sexual nudity

3. Violence

Violent content includes descriptions, references, depictions and encouragement of fighting, bloody / gory scenes, animal cruelty, crude content. Violent content can range from less extreme to graphic content.

a) The content is likely to incite violence against individuals
    - Including but not limited to, based on protected characteristics such as race, religion, sex, sexual orientation, disability or gender reassignment
b) The content depicts violence against individuals
    - Including but not limited to, based on protected characteristics such as race, religion, sex, sexual orientation, disability or gender reassignment
c) The content includes threats to kill or of violence against an individual or individuals
d) The content depicts, promotes or encourages fighting
e) The content depicts, promotes or encourages warfare
f) The content is of a gory and bloody nature
g) The content promotes Female Genital Mutilation (FGM)
h) The content amounts to an offer to sell or supply a firearm or other prohibited or controlled weapon (including where it would be illegal for a child to buy such an item)
i) The content is likely to encourage or promotes use of a firearm or other type of controlled weapon (including where it might encourage obtaining or carrying of a knife by a child)

4. Dangerous and risky behaviour

Content that describes, references, depicts and encourages behaviours that are age inappropriate and may cause danger to the user or others

a) The content appears to depict, promote or encourage alcohol consumption
b) The content is likely to encourage or promotes the taking of an illegal drug or psychoactive substance
c) The content appears to depict, promote or encourage dares / challenges involving potential bodily harm to users or others
d) The content appears to depict, promote or encourage dares / challenges involving potential financial loss
e) The content appears to be a scam
f) The content amounts to an offer to sell or supply an illegal drug or psychoactive substance

5. Sexual

Sexual material related to / aimed at children – material that encourages or depicts engagement in activity of a sexual nature by children

Adult sexual material – material that contains images and/or language of a strong sexual nature which is shared for the primary purpose of sexual arousal or stimulation

Sexually suggestive material – material that contains references / images and languages of a mild sexual nature, not necessarily broadcast for the purpose of sexual arousal or stimulation

Sexual material related to / aimed at children

a) The content amounts to incitement of a child to engage in sexual activity
b) The content appears to promote, encourage or depict sexting of indecent images by under 18s (creating, possessing, copying or distributing indecent or sexual images of children and young people under the age of 18)

Adult sexual material

c) The content depicts nudity in a sexual context
d) The content includes erotic literature
e) The content appears to include explicit fanfiction / imaginative sexual scenarios
f) The content is pornographic (e.g. it appears to have been produced solely or principally for the purposes of sexual arousal)
    - This can include extreme pornography
    - This can include revenge pornography

Sexually suggestive material

g) The content depicts suggestive poses / behaviours
h) The content depicts sexually suggestive scenes from films
i) The content appears to include non-explicit but suggestive fanfiction / imaginative scenarios
j) The content appears to include ‘dirty’ jokes and comments

6. Intimidation, aggression and hate speech

Content that describes, references, depicts and encourages intimidation and harm towards others

a) The content appears to amount to bullying or harassment of an individual or individuals
    - Including cyberbullying and trolling
b) The content appears likely to incite hatred against people based on protected characteristics (race, religion, sex, sexual orientation, disability or gender reassignment)
c) The content is abusive and targets protected characteristics (race, religion, sex, sexual orientation, disability or gender reassignment)
d) The content promotes or supports a proscribed terrorist organisation
e) The content promotes or supports organised immigration crime
f) The content promotes or supports modern slavery
g) The content promotes or encourages coercive behaviour
h) The content includes swearing / foul language aimed at an individual or group
i) The content appears to share or promote personal details about an individual (i.e. doxing)

7. Information / disinformation

Content that describes and depicts information about the world and other people

a) The content appears to promote or spread fake news
b) The content appears to include information that is likely to be harmful to the child's health