RESPONSE ON BEHALF OF WHATSAPP IRELAND LIMITED TO OFCOM'S ONLINE SAFETY CALL FOR EVIDENCE (PHASE 2) [MARCH 2023]

We would like to thank Ofcom for the opportunity to provide information in response to this call for evidence.

WhatsApp Ireland Limited[1] ("**WhatsApp**") has supported the UK Government's development of the Online Harms framework and shares the UK Government's stated policy objectives, to make the internet safer while protecting the vast social and economic benefits it brings to billions of people each day.

[REDACTED] At the heart of these concerns is the uncertain impact of the Bill on users' privacy and security due to a lack of clarity on its impact to private messaging.

Today end-to-end encryption (E2EE) protects users' privacy and security from risks like hacking and cyber crime which are on the rise[2]. Indeed, encryption is vital to ensuring the safety and security of vulnerable groups - including children - in their everyday lives.[3] The UN recently called on states to avoid weakening encryption through proposals for backdoors or proactive technologies like client-side scanning[4].

Experts have repeatedly stated that scanning technologies like 'client side scanning' (which Ofcom may have powers to impose under the latest Bill) would undermine E2EE.[5] Put simply, it would be like installing a camera in everyone's living room, with an algorithm deciding whether footage should be sent to the authorities or not. Experts agree that in addition to privacy concerns, such technology would be ineffective and would open the door to broader abuses by hackers and hostile states.[6]

While we welcome reassurance from the UK Government about their support for E2EE - for example, Paul Scully's statement in the House of Commons that *"the powers do not represent a ban on or seek to undermine any specific type of technology or design, such as end-to-end encryption"*[7] - we believe the Bill could be clearer about its applicability to private messaging.

Our three main concerns with the regime in respect of private messaging relate to:

[1] WhatsApp Ireland Limited is the current service provider for UK users. In April 2023, the service provider for UK users will be changing to WhatsApp LLC.

[2] The Crime Survey for England and Wales found 3.7 million reported incidents in 2019-20 of members of the public being targeted by credit card, identity and cyber-fraud, making it the crime UK citizens are most likely to fall victim to.

[3] Unicef Innocenti: Encryption, Privacy and Children's Right to Protection from Harm, October 2020; The Guardian: End-to-end encryption protects children, say UK information watchdog, January 2022; CRIN: Privacy and Protection: A children's rights approach to encryption

[4] UN Human Rights Office of the High Commissioner: The right to privacy in the digital age, September 2022

[5] Bugs in our Pockets: The Risks of Client-Side Scanning

[6] EFF: Why Adding Client-Side Scanning Breaks End-To-End Encryption; De Montoye et al: Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning

[7] https://hansard.parliament.uk/commons/2022-12-05/debates/E155684B-DEB0-43B4-BC76-BF53FEE8086A/OnlineSafetyBill

(i) technology notices and powers which could require providers of private messaging to put in place widespread scanning and surveillance of users' private conversations which would be incompatible with end-to-end encryption ('E2EE');

(ii) category 1 obligations being applied to private messaging that are incompatible with E2EE; and

(iii) how the risk assessment process (which appears to be centred around content moderation obligations), will work for private messaging where content monitoring and moderation would be inconsistent with user expectations of private messaging as well as incompatible with the operation of an E2EE service.

Importance of a differentiated approach

WhatsApp supports proportionate and workable regulation for private messaging. We look forward to continued engagement as Ofcom develops and subsequently consults on a differentiated approach to regulating private messaging - including where users are choosing to use E2EE services. This would involve designing a code of practice and risk assessment approach that is appropriate for services that do not host or have access to the contents of people's private messages.

In particular, it is important that the resulting code of practice and risk assessment processes does not require private messaging services to monitor and / or moderate the content of people's private conversations. This would not be in line with user expectations around the privacy of their personal communications. For example, the European Commission's 2016 public consultation on the ePrivacy rules in the EU found that, of the 27,000 respondents surveyed "*nine in ten agree they should be able to encrypt their messages and calls, so they can only be read by the recipient.*"[8]

Safety and integrity in private conversations

There are however approaches to user safety that do not involve content monitoring or moderation. [REDACTED], as an E2EE private messaging service, WhatsApp does not have access to the content of conversations on the service[9]. In order to protect our users from harm on WhatsApp, while remaining world-class on user privacy, we strive to design the environment to make harm less likely to happen in the first place, and to empower people to keep themselves safe if they do encounter harm. We consider a wide range of harms – including risks like cyber security and hacking.

Our approach to thinking about the risk of harm is therefore based on:

- Preventing abuse through product design, stopping it from happening in the first place through product functionality and features.
- A strong suite of user controls allowing people to control their experience and keep themselves

[8] https://digital-strategy.ec.europa.eu/pl/node/4609/printable/pdf
[9] Unless reported to us by users themselves, for more information please see here: https://faq.whatsapp.com/408155796838822/?helpref=faq_content

safe on the service.

- Designing ways to catch those who, despite our best efforts, violate our policies or use our service to cause harm.

- Collaboration with outside experts on everything from safety measures to education campaigns, and working with law enforcement, including to respond to requests based on applicable law and policy.

To help keep users safe, WhatsApp has developed and deployed a number of advanced tools to prevent unwanted contact via the platform.[10] The service also has design features such as messaging forwarding limits, prominent opportunities to block contacts, and a number of privacy settings to control who can see users' personal information (such as their profile photo or status), and how users can be added to groups – all of which are designed to prevent abuse for users of all ages. These measures are all content-agnostic.

In addition, WhatsApp requires a phone number to sign up to the service and provides no on-platform ability to search for unconnected individual accounts. WhatsApp also provides a number of entry points to report problematic accounts or content.

In short - our focus is on architecting the product to limit all users' exposure to harm and detecting and responding to harm based on the information available to us. Privacy and data minimisation are at the heart of this approach.

We are happy to work with Ofcom to identify additional information that may help to inform Ofcom's thinking. We look forward to continuing our constructive relationship.

[REDACTED]

Q1. To assist us in categorising responses, please provide a description of your organisation, service or interest in protection of children online.

[REDACTED]

Q2. Can you identify factors which might indicate that a service is likely to attract child users?

In particular, please provide evidence explaining:

- *the types of services which are likely to attract child users;*
- *any functionalities or other features of a service which are particularly likely to attract child users;*
- *the type of content that is likely to attract child users;*
- *if or how the factors you've identified may differ depending on the age of a child user; and*

[10] See further here: https://faq.whatsapp.com/1104252539917581/?helpref=uf_share

- *whether there are any noticeable patterns in the activity of child users. Where possible, please specify whether this evidence relates to child users in the UK or globally*

Q3. What information do services have about the age of users on different platforms (including children)?

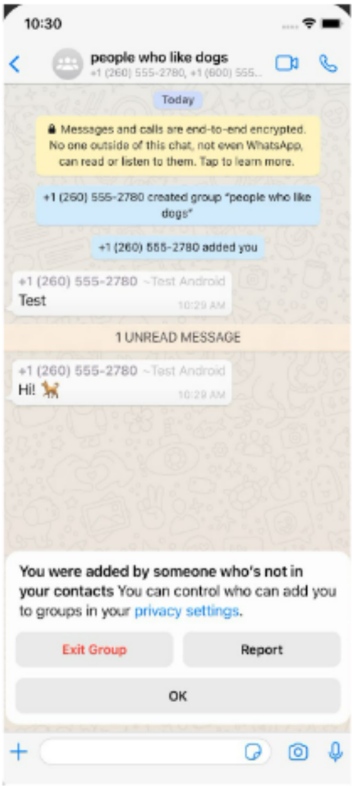
In particular, please provide evidence explaining:

- *the methods used to gather any information that can assist in estimating or assuring a user's age, either at the point a user first accesses the service or subsequently;*
- *what, if any, mechanisms are available to enable services to identify children in different age groups (for example children below age 13, aged 13-15, or aged 15-17); and*
- *how approaches to assessing the age of users are evolving.*

[**Answering 2 and 3]**

[REDACTED]

Finally, WhatsApp offers a robust set of controls and defaults for all users on the platform today. We make controls available to users proactively, empowering users to create the right WhatsApp experience for themselves.



Messages and calls are end-to-end encrypted. No one outside of this chat, not even WhatsApp, can read or listen to them. Tap to learn more.

+1 (260) 555-2780 created group "people who like dogs"

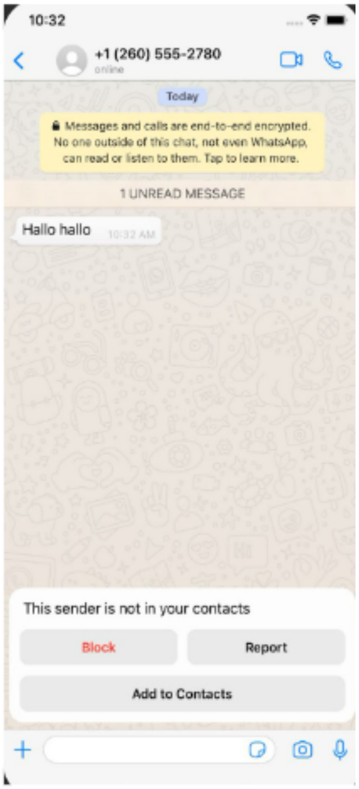
+1 (260) 555-2780 added you

+1 (260) 555-2780 ~Test Android
Test 10:29 AM

1 UNREAD MESSAGE

+1 (260) 555-2780 ~Test Android
Hi! 🐕 10:29 AM

You were added by someone who's not in your contacts You can control who can add you to groups in your privacy settings.



Messages and calls are end-to-end encrypted. No one outside of this chat, not even WhatsApp, can read or listen to them. Tap to learn more.

1 UNREAD MESSAGE

Hallo hallo 10:32 AM

This sender is not in your contacts

Q4. How can services ensure that children cannot access a service, or a part of it?

In particular, please provide evidence explaining:

- *how age assurance policies have been developed to date and what age group(s) they are intended to protect;*
- *if the service is tailored to meet age-appropriate needs (for example, by restricting specific content to specific users or part of a service), how this currently works or could work;*
- *how the efficacy of age assurance policies is or could be monitored; and*
- *how services can identify users that do not meet any relevant age limits and how is this appropriately addressed?*

In order to create a WhatsApp account, the individual must agree to the Terms of Service. The Terms of Service provide: "*If you live in a country in the European Region, you must be at least 16 years old to use our Services or such greater age required in your country to register for or use our Services.*"

Creating an underage account is a violation of the Terms of Service, as is registering an account on behalf of someone under the age of 16 years old for users in the UK and European Region. WhatsApp disables accounts in response to reasonably verifiable reports of an account being created in violation of the Terms of Service.[11]

The WhatsApp Help Centre gives parents / guardians information on how to delete an account if it has been created by an individual under the age of 16 years old. Parents / guardians can also report underage accounts directly to WhatsApp via email.[12]

WhatsApp has established a dedicated Trust & Safety team that reviews reports regarding users that are allegedly below the age of 16 years old. If it is reasonably verifiable that the account belongs to an individual under the age of 16 years old, WhatsApp will disable the account. This reporting mechanism is highly visible through WhatsApp's dedicated FAQ page on minimum age, which directly links to the channel to make a report.[13]

Q5. What age assurance and age verification or related technologies are currently available to platforms to protect children from harmful content, and what is the impact and cost of using them?

In particular, please provide evidence explaining:

- *how these technologies can be assessed for effectiveness or impact on users' safety;*
- *how accurate these technologies are in verifying the age of users, whether accuracy varies based on any user characteristics, and how effective they are at preventing children from accessing harmful content;*
- *any potential unintended consequences of implementing age assurance (such as risk of bias or exclusion), and how these can be mitigated;*

[11] See more at: https://faq.whatsapp.com/general/security-and-privacy/minimum-age-to-use-whatsapp/?lang=en

[12] See more at: https://faq.whatsapp.com/general/security-and-privacy/minimum-age-to-use-whatsapp/?lang=en

[13] *Ibid*.

- *the safeguards necessary to ensure users' privacy and access to information is protected, and over restriction is avoided;*
- *which methods of age assurance users prefer, when offered a number of ways to verify their age;*
- *the cost of implementing and operating such technologies;*
- *how age assurance and age verification or related technologies may be circumvented; and*
- *what mitigations exist to reduce circumvention among users.*

[REDACTED]

The effective verification of age in an online environment remains a developing and technically challenging area, particularly in the context of services such as the WhatsApp Service which collects limited categories of data and utilises end-to-end encryption. Figuring out how to verify age sufficiently accurately without violating data minimisation principles and/or excluding significant disproportionate numbers of users (both young and old) from online services who are unable to prove their age is a challenge that the entire industry is facing.

[REDACTED]

Q6. Can you provide any evidence relating to the presence of content that is harmful to children on user-to-user and search services?

We are interested in evidence about the quantity or presence of such content on services of particular types or on user-to-user and search services in general. This could include, for example, the findings of relevant investigations, transparency reports and research papers that demonstrate how such content might vary across different services or types of service, or across services with particular groups of users, features or functionalities.

In particular, please provide evidence explaining:

- *which groups or ages of children, if any, are more likely to encounter content that is harmful to children;*
- *the prevalence of primary priority content on user-to-user and search services; • the prevalence of priority content on user-to-user and search services; and*
- *the types of service children are most likely to encounter different types of harmful content on*

Q7. Can you provide any evidence relating to the impact on children from accessing content that is harmful to them?

In particular, please provide evidence explaining:

- *the impact of harm on children who have encountered primary priority content;*
- *the impact of harm on children who have encountered priority content;*
- *any specific risks children may encounter on search services associated with harmful content; and*
- *any differences in the impact of priority content on children in different age groups.*

[Answering 6 and 7]

As an end-to-end encrypted private messaging service, WhatsApp does not have access to the content of conversations on the service unless reported to us by users themselves.[14]

In order to protect our users from harm on WhatsApp, while remaining world-class on user privacy, we design the environment to make harm less likely to happen in the first place, and to help people keep themselves safe if they do encounter harm – illegal or otherwise. In short, we seek to design the service in a way that makes it difficult for bad use cases to thrive.

Q8. How do services currently assess the risk of harm to children in the UK from content that is harmful to them?

In particular, please provide evidence explaining:

- *how risks from harmful content are identified (including any relevant internal processes, policies and documents); and*
- *in considering the potential risk that children may encounter harmful content, the extent to which services factor in evidence on users' behaviour and age.*

Q9. What are the exacerbating risk factors services do or should consider which may have an impact on the risk of harm to children in the UK?

In particular, please provide evidence of:

- *how the user base of the service may have an impact on the risk of harm to children;*
- *how the business model of the service may have an impact on the risk of harm to children;*
- *the functionalities or features of services which may have an impact on the risk of harm to children; and*
- *what mitigations exist for these risk factors.*

Q10. What are the governance, accountability and decision-making structures for child user and platform safety?

As part of your answer, please outline how different teams may consider child user safety risks across different business functions such as product development, management, engineering, public policy, safety, legal, business development and marketing.

[14] For more information, please see here: https://faq.whatsapp.com/408155796838822/?helpref=faq_content

Please consider:

- *how staff are/should be trained in a service to understand how their own roles and responsibilities can create risks to child user safety;*
- *how services can ensure consistency in consideration of child user safety across teams;*
- *examples of best practice or innovative approaches to governance, accountability and decision making; and*
- *possible costs associated with assessing risks of harm to child users, including specific reference to costs associated with ensuring services have governance and decision-making processes for child user safety where applicable.*

[Answering 8,9 and 10]

As we set out above, in order to protect our users from harm on WhatsApp, while remaining world-class on user privacy, we strive to design the environment to make harm less likely to happen in the first place, and to empower people to keep themselves safe if they do encounter harm. This means considering a wide range of harms users face – including risks like cyber security and hacking.

As we've explained, WhatsApp is an end-to-end encrypted private messaging service and therefore does not have access to the content of conversations on the service (unless reported to us by users themselves, as discussed above).

Within this context, safety and integrity risks are considered by WhatsApp at various stages in the product development process, including in product roadmapping and the design and build process. In addition to this, we also have dedicated teams that are focused solely on building safety and integrity features within the product.

We have previously presented WhatsApp's safety and integrity approach to Ofcom, and would be happy to engage in further discussion in respect of our approach, to help illustrate how private messaging platforms can deal with illegal and harmful conduct.

Q11. What can providers of online services do to enhance the clarity and accessibility of terms of service and public policy statements for children (including children of different ages)?

Please submit evidence about what approaches make terms or policies clear and accessible to children.

WhatsApp makes its terms of service and policies available online to everyone[15], to ensure that they are easily accessible and that users are clear on the terms they sign up to. These terms can also be easily accessed from within the Settings menu in the app. In addition to the terms of service, we also publish a range of Help Centre articles on our website that provide guidance to our users in simple and clear

[15] WhatsApp terms of service and privacy policy: https://www.whatsapp.com/legal/terms-of-service/?lang=en
WhatsApp Help Centre article 'How to Use WhatsApp responsibility" https://faq.whatsapp.com/1325842477576427/?locale=en_US and related video https://www.youtube.com/watch?v=-a4um6gYPKY

terms. See for example information on "How to use WhatsApp responsibly" that is available both as a Help Centre article and as a video.

Q12. How do terms of service or public policy statements treat 'primary priority' and 'priority' harmful content?

Please outline as part of your answer:

- *what services currently cover in their terms of service and public policy documents in relation to primary priority and priority harmful content and why, including any reference to what is considered harmful content, any measures to identify it, and sanctions applied to users who are in breach;*
- *whether when drafting these documents, the specific needs of children are considered;*
- *evidence of the process, time and any costs involved in developing these terms; and*
- *whether you have any evidence about how child users engage with terms of service or public policy statements, or whether children understand what they mean in practice.*

Q13. What can providers of online services do to enhance children's accessibility and awareness of reporting and complaints mechanisms?

Please submit evidence about what features make user reporting and complaints systems accessible to children, considering:

- *reporting or complaints routes for registered and non-registered users; • how services could encourage children to report content;*
- *how to ensure that reporting and complaints mechanisms are not misused;*
- *the key choices and factors involved in designing these mechanisms;*
- *how to ensure that reporting and complaints mechanisms are user-friendly and accessible to children;*
- *whether particular consideration is given to the different needs of child users, for example children of different ages;*
- *whether user reports are anonymous to the service;*
- *whether users are notified that their reports are anonymous to other users; and*
- *what happens to users who have their content or account reported.*

Q14. Can you provide any evidence or information about the best practices for accurate reporting and/or complaints mechanisms in place for legal content that is harmful to children, or users who post this content, and how these processes are designed and maintained?

We are interested in obtaining evidence on:

- *how users, including children, report harmful content on services (including the mechanisms' location and prominence for users, and any screenshots you could provide) and whether this is or ought to be separate to complaints procedures;*
- *whether users need to create accounts to access reporting and complaints mechanisms;*

- *what type of content or conduct, users and non-users may make a complaint about or report, including any specific lists or categories;*
- *whether reporting and complaints mechanisms are effective, in terms of identifying content that is harmful to children, and how to determine effectiveness;*
- *whether there are any reporting or complaints mechanisms you consider to be less effective in terms of identifying content that is harmful to children, and how you determine this;*
- *the use of trusted flaggers (and if reports from trusted flaggers should be prioritised over reports or complaints from users); and*
- *the cost involved in designing and maintaining reporting and/ or complaints mechanisms.*

[Answering 13 and 14]

We encourage users to report problematic messages or accounts to us. Users can report other accounts by tapping the contact name > tapping Report Contact > and then tapping Report And Block (for iOS) or by tapping More options > tapping more > and then tapping Report (for Android). Users can follow similar steps to report a group instead of an account. When a user reports another account or group, WhatsApp receives the last five messages sent to the reporting user by the reported user or group. WhatsApp also receives the reported group or user ID, information on when the message was sent, and the type of message sent (image, video, text, etc.). Alternatively, users can also long press an individual message and tap on the overflow menu to report that particular message to WhatsApp.

Alongside these reporting options, users receive an additional option to either block the contact or exit the group, as applicable. No matter which option the user chooses, the reported account or the group is not notified of the reporter's actions.

Whenever a user receives a message from someone outside of their address book, we display a message asking if the user wants to “block” or “report” the contact. Users can also control who can add them to a group, and if they are added to a group by someone outside of their address book, we display the same dialog giving the user an easy way to report the individual and exit the group. While we have had this system dialog for a number of years, we are always looking for ways to improve the reporting experience for users. Last year, for example, we made improvements to this dialog to make it more prominent so that it persistently remains visible on the chat thread unless the user dismisses it or responds to a message they received from the non-contact.

These reporting tools are available only to users, since receiving unwanted messages or contacts requires users to own a phone number and a WhatsApp account. More details on blocking and reporting on WhatsApp are available here.[16] The Help Centre provides information on how to block and report on iPhone, Android, and KaiOS.

Q15. What actions do or should services take in response to reports or complaints about online content harmful to children (including complaints from children)?

Please provide relevant evidence explaining your response to this question.

[16] https://faq.whatsapp.com/408155796838822/?helpref=hc_fnav

[REDACTED]

Q16. What functionalities or features currently exist that are designed to prevent or mitigate the risk or impact of content that is harmful to children?

In particular, please provide evidence explaining:

- *any functionalities or features available to services, which you consider can effectively prevent harm to children; and*
- *any functionalities or features in development that services could consider implementing to mitigate the risk or impact on children of content harmful to them.*

Q17. To what extent does or can a service adopt functionalities or features, designed to mitigate the risk or impact of content that is harmful to children on that service?

In particular, please provide evidence explaining:

- *what control can or should children have over what they are shown or content that is delivered to them;*
- *to what extent the features or functionalities identified are reliant on other technology – for instance, age assurance, age verification or ID verification mechanisms;*
- *the costs or cost drivers involved in developing these features or functionalities;*
- *whether child safety is incorporated into the product design and development processes;*
- *to what extent evidence is considered about child user behaviour when developing features or functionalities intended to enhance user safety;*
- *what evidence can be provided relating to measures, including evidence around the impact and effectiveness of these techniques, in terms of reducing harm to children; and*
- *how services assess the impact of potential mitigations on users' privacy and freedom of expression and minimise the risk of over restriction.*

Q18. How can services support the safety and wellbeing of UK child users as regards to content that is harmful to them?

In particular, please provide evidence explaining:

- *whether this involves or should involve support provided through the platform (e.g. signposting to resources);*
- *whether this involves or should involve off-platform support (e.g. funding or facilitating programmes);*
- *how are these interventions or should these interventions be embedded into the user journey via service design;*
- *how effective these types of interventions, in terms of minimising harm from and impact of harmful content to children are; and*
- *the costs involved in implementing the support measures you have described.*

[Answering 16, 17 and 18]

Please see our response to Question 4 above, where we describe how we design the WhatsApp environment to make harm less likely to happen in the first place, and to help people keep themselves safe if they do encounter harm.

WhatsApp has implemented a suite of in-app tools and features to help all users have a private, safe and secure forum to communicate. In particular, WhatsApp has implemented the following tools and features:

- **Controls on contacting users.** A user must have another user's phone number to contact them. A user cannot 'search' or contact someone on the Service unless they have their phone number. Equally, the Service does not 'suggest' other users that a user may want to connect with in order to expand their WhatsApp contacts list. This limits the exposure of users to unwanted and unsolicited contact.

- **Safety tools to mute, block or report unwanted interactions.** The Service has easy-to-use in-app tools that allow users to restrict unwanted interactions from those people that do have their number. For example, if a user receives a message from someone who is not saved in their WhatsApp contact list, the Service immediately asks if the user would like to "block" or "report" that other user. A user can also block another user at any time from within a chat or by selecting individual contacts to be added to their blocked contacts list. The chosen contact will not be notified that they have been blocked. The "mute notifications" function also allows users to easily stop receiving notifications from chosen contacts or groups without those contacts or groups being notified that their notifications have been muted. The Service also makes it easy for users to report problematic contacts or messages from within the app. When the sender or recipient of a message reports it, the content is sent for review. WhatsApp takes appropriate action against activity that is in violation of WhatsApp's safety and security policies, including potentially banning a user's phone number thereby disabling a user's account. Group privacy settings also allow users the ability to control who can add them to a group.[17] For example, users can limit who can add them to groups to only their contacts or to exclude users even within their own contacts list.[18] All users, including teens, have additional control over features such as push notifications, alerts, and updates through their device settings.

- **Limited publication and sharing capabilities.** Users are not required to provide a profile picture or short biography in the "About" section of the user profile. If users choose to include a profile photo and "about" information, these are only visible to other users who have that user's phone number or are in a group with that user. Users can also disable access to their profile photo or About information, so that it is not visible to any other user.[19] WhatsApp has

[17] See more at: https://faq.whatsapp.com/general/security-and-privacy/how-to-change-group-privacy-settings

[18] See more at https://faq.whatsapp.com/general/security-and-privacy/how-to-change-group-privacy-settings

[19] See more at: https://faq.whatsapp.com/general/security-and-privacy/how-to-change-your-privacy-settings/ and https://faq.whatsapp.com/general/contacts/cant-see-a-contacts-profile-information

published easy-to-follow guidelines on how users can control who can see their profile photo,[20] "last seen" information,[21] About information,[22] status updates[23] and read receipts.[24] WhatsApp has also published a detailed guide for users on how to stay safe when sharing personal information.[25]

- **Robust security measures.** To protect privacy and security, all personal messages sent using the Service are secured with end-to-end encryption.[26] As users' messages are encrypted, only the message sender and the message recipient have the encryption key needed to unlock and read the message. The Signal Protocol, designed by Open Whisper Systems, is the basis for the Service's end-to-end encryption. This industry-leading end-to-end encryption protocol is designed to prevent third parties and WhatsApp from having plaintext access to messages, including all its contents, or calls. What's more, even if encryption keys from a user's device are ever physically compromised, the compromised encryption keys cannot be used to go back in time to decrypt previously transmitted messages. In addition to the encryption technology used to safeguard messages sent using the Service, users can implement further security measures to safeguard their account through the Service's two-step verification feature. When enabling this feature, users create a unique PIN that adds more security to their WhatsApp account.[27] Users also have the option to enable TouchID, FaceID or Android fingerprint lock to secure their WhatsApp accounts.[28]

- **Product changes to help limit the spread of harmful content, including misinformation, and strengthen message privacy.**

    - On the Service, a user can only forward a message to up to five chats at one time, making WhatsApp one of the few services to intentionally limit sharing. Messages which are labelled as *'Forwarded many times'* can only be forwarded to one other chat at a time[29]. Although highly forwarded messages make up a very small percentage of all messages sent on the Service, the introduction of this limit has further reduced the number of these messages by over 70% globally.

[20] See more at: https://faq.whatsapp.com/android/account-and-profile/how-to-edit-your-profile/?lang=en

[21] A user's "last seen" information tells other users the last time that user used WhatsApp, or if they are online (see: https://faq.whatsapp.com/general/chats/about-last-seen-and-online).

[22] See more at: https://faq.whatsapp.com/iphone/account-and-profile/how-to-edit-your-profile

[23] See more at: https://faq.whatsapp.com/general/status/about-status-privacy

[24] See more at: https://faq.whatsapp.com/general/security-and-privacy/how-to-change-your-privacy-settings

[25] See more at: https://faq.whatsapp.com/general/security-and-privacy/staying-safe-on-whatsapp/

[26] For further detail on WhatsApp use of end-to-end encryption, please see the WhatsApp Security White Paper available at: https://www.whatsapp.com/security/WhatsApp-Security-Whitepaper.pdf

[27] See more at: https://faq.whatsapp.com/general/security-and-privacy/account-security-tips and https://faq.whatsapp.com/general/verification/about-two-step-verification?category=5245245

[28] See more at: https://faq.whatsapp.com/iphone/security-and-privacy/how-to-usetouch-id-or-face-id-for-whatsapp and https://faq.whatsapp.com/android/security-and-privacy/how-to-use-android-fingerprint-lock

[29] See more at: https://faq.whatsapp.com/general/chats/about-forwarding-limits

  - ○ Further, users also have the option to send messages that will disappear.[30] Once this feature is enabled, the user's messages will disappear after their chosen time period (seven days is the default, but users' can also choose 24 hours or 90 days). The purpose of these controls is to replicate in-person communication as much as is possible and to allow individuals to exercise control over their communications.

- **Educational resources.** All users, including those under 18, are provided with detailed and easy-to-follow guidance on how to use the Service and protect their privacy and security. For example, WhatsApp's resources facilitate users in learning how to customise privacy settings for both individual contacts and group chats.[31] WhatsApp also emphasises the importance of responsible use of the Service by for example providing specific information on best practice and practices to avoid when using the Service.[32] Additionally, following the step-by-step guide, individuals can find out how to block specific contacts from interacting with specified contacts.[33] Further, as described above, if an individual under the age of 16 years old has created an account, an easy to follow guide tells individuals how they can delete such an account.[34] These accounts can also be reported. The information required in such a scenario is outlined in a detailed FAQ for parents / guardians, along with advice on how to protect personal details when making a report.[35] Once it is verified that an account belongs to an individual under the age of 16, it will be disabled / banned.

In addition to the above educational materials, WhatsApp has a dedicated safety page in its Help Centre[36], which includes information on:

- **Controls over what users share:** WhatsApp encourages users to think carefully before they decide to share anything.[37]
- **Banning of accounts:** accounts may be banned if they have violated the Terms of Service.[38]
- **Spam and hoax messaging:** a guide is provided for users on the key hallmarks to identify spam or hoax messages.[39]
- **Mental health resources:** links to mental health providers are listed if a user needs support.[40]

[REDACTED]

[30] See more at: https://faq.whatsapp.com/general/chats/about-disappearing-messages

[31] See more at: https://faq.whatsapp.com/general/security-and-privacy/how-to-change-group-privacy-settings/; https://faq.whatsapp.com/general/security-and-privacy/how-to-change-your-privacy-settings/

[32] See more at: https://faq.whatsapp.com/general/security-and-privacy/how-to-use-whatsapp-responsibly/?lang=en

[33] See more at: https://faq.whatsapp.com/iphone/security-and-privacy/how-to-block-and-unblock-contacts

[34] See more at: https://faq.whatsapp.com/android/account-and-profile/how-to-delete-your-account/?lang=en

[35] See more at: https://faq.whatsapp.com/general/security-and-privacy/minimum-age-to-use-whatsapp/

[36] See more at: https://faq.whatsapp.com/general/security-and-privacy/staying-safe-on-whatsapp

[37] See more at: https://faq.whatsapp.com/general/security-and-privacy/staying-safe-on-whatsapp

[38] See more at https://faq.whatsapp.com/general/account-and-profile/about-account-bans

[39] See more at https://faq.whatsapp.com/general/security-and-privacy/about-spam-and-unwanted-messages

[40] See more at: https://faq.whatsapp.com/general/security-and-privacy/global-suicide-hotline-resources/

Q19. With reference to content that is harmful to children, how can a service mitigate any risks to children posed by the design of algorithms that support the function of the service (e.g. search engines, or social and content recommender systems)?

In particular, please provide evidence explaining:

- *if different from the risk assessment process outlined in response to Q8 how services assess the risk to children from algorithms central to the function of the service;*
- *what safeguards services have in place to mitigate the risks posed by algorithms (e.g. testing them before they are put into use, and monitoring their performance in real world settings);*
- *what safeguards services have in place to mitigate the risks posed using recommender systems in particular (e.g. providing users with controls over what they are shown, such as through keyword filters);*
- *additional requirements for safety in algorithms (e.g. accurate content categorisation); • the costs involved in implementing these safeguards. In the absence of specific costs, please provide indication of the key cost drivers;*
- *how services can measure the effectiveness of these safeguards, in terms of reducing harm to users;*
- *what information services can provide to demonstrate the effectiveness of such safeguards; and*
- *how services can assess the impact of these safeguards on users' privacy and minimise the risk of over restriction.*

[N/A RESPONSE]

Q20. Could improvements be made to content moderation to deliver greater protection for children, without unduly restricting user activity?

If so, what? In particular, please provide relevant evidence explaining:

- *improvements in terms of user safety and user rights (e.g. freedom of expression), as well as any relevant considerations around potential costs or cost drivers;*
- *evidence of the effectiveness of existing moderation systems; and*
- *examples of where relevant content moderation processes are particularly good or poor.*

[N/A]

Q21. What automated, or partially automated, moderation systems are currently available (or in development) for content that is harmful to children?

In particular, please consider:

- *the suitability of automated (or part-automated) moderation systems to identify content that is harmful to children;*

- *whether and how automated moderation systems differ by the type of content (e.g. text, image or video);*
- *how in-house automated content moderation systems are developed and (in the case of technology which uses AI or machine-learning) trained or tested;*
- *how the data used to develop, train, test or operate content moderation systems is sourced and whether it is representative of the intended real-world scenario;*
- *the range or quality of third-party content moderation system providers available in the UK;*
- *how effective automated content moderation systems are, in terms of identifying target content that is harmful to child users, and how this may vary by harm;*
- *what evidence is available to assess the accuracy of automated moderation techniques (e.g. regarding the frequency of false positives/negatives);*
- *how action is taken in relation to content identified by automated means as potentially being harmful to children (e.g. automated action, human action, or further review);*
- *what safeguards are employed to mitigate adverse impacts of automated content moderation, e.g. on privacy and/or freedom of expression;*
- *whether certain types of automated moderation techniques might be better suited to certain harms or types of content and why; and*
- *what barriers and costs are involved in deploying these automated moderation systems.*

[N/A]

Q22. How are human moderators used to identify and assess content that is harmful to children?

In particular, please consider:

- *the typical role of an effective human moderator;*
- *how to determine the level of human moderation required by a platform, including by type of harmful content;*
- *whether moderators are employed by the service, outsourced, or volunteers; • whether moderators are vetted, and how; and*
- *the type of coverage (e.g. weekends or overnight, UK time) moderators provide*

Q23. What training and support is or should be provided to moderators?

In particular, please consider:

- *whether certain moderators are specialised in certain harms or speech issues;*
- *whether training is provided or updated, and frequency of these; and*
- *whether moderators are trained to identify content that is harmful to children.*

Q24. How do human moderators and automated systems work together, and what is their relative scale? How should services guard against automation bias?

In particular, please provide evidence explaining:

- *the costs involved in these practices. In the absence of specific costs, please provide indication of cost drivers (e.g. moderator location) and other relevant figures (e.g. number of moderators employed, how many items the platform moderates per day); and*
- *how services can assess the accuracy and consistency of human moderation teams.*

[Answering 22, 23, and 24]

[REDACTED]

Q25. In what instances is content that is harmful to children, that is in contravention of terms and conditions, removed from a service or the part of a service that children can access?

Please outline the circumstances in which content that is harmful to children is removed and how this may differ by type of harmful content.

As we explain above, as an end-to-end encrypted private messaging service, WhatsApp's approach to integrity focuses on product design and account-level actions to address abuse and not on content moderation of individual pieces of content.

Q26. What other mitigations do services currently have to protect children from harmful content?

In particular, please provide evidence explaining:

- *available mitigations to prevent children from accessing primary priority content;*
- *available mitigations to protect children from accessing priority content;*
- *what types of service (e.g. social media, search, gaming etc.) available mitigations are suitable for;*
- *the costs of implementing mitigations;*
- *if applicable, the potential risks associated with mitigations;*
- *if applicable, any non-cost barriers of implementing mitigations;*
- *how effective the existing mitigations platforms have in place are;*
- *how 'other features' (alongside design of functionalities and algorithms) are designed to ensure safety to children; and*
- *what, if any, mitigations are feasible but may not be currently available on services likely to attract children.*

In addition to designing our product to be safe, we also deploy content-agnostic measures to further reduce potential exposure to harm.

For example, when a user is contacted by someone who is not already a saved contact in their address book, we provide a prompt asking users if they would like to block or report the individual. And unless the recipient chooses to engage in a conversation, images sent by a user outside of the address book are blurred by default, and links sent by them are delivered as plaintext rather than in the form of a clickable link.

In addition as noted above, we have implemented forward labels to educate users that a message they have received has been previously forwarded to them from another person or group, as well as forward limits to help limit message virality. The introduction of forward limits resulted in a 70% reduction of highly forwarded messages globally.

Q27. Where children attempt to circumvent mitigations in place on a service, what further systems and processes can a service put in place to protect children?

Here we are particularly interested in the circumvention of measures that are not age assurance technologies, which we ask about in Q5. In particular, please provide evidence explaining:

- *the ways in which some child users may attempt to circumvent mitigations put in place to protect them from harmful content;*
- *the ways in which services can combat these, particularly in relation to primary priority content, which children should be prevented from accessing; and*
- *examples of best practice for how to prevent children from encountering primary priority content.*

[N/A]

Q28. Other than those covered above in this document, are you aware of other measures available for mitigating the risk, and impact of, harm from content that is harmful to children?

We would be interested in any evidence you can provide on their efficacy, in terms of reducing harm to child users, cost and impact on user rights and user experience.

In relation to private messaging services, we are aware of claims that there are approaches to scanning private message contents to identify content like CSAM without compromising E2EE, such as client-side scanning. However, cryptographers and privacy and computer security experts agree that these approaches do in fact compromise E2EE and undermine people's privacy and security.[41] This is because the fundamental premise of E2EE is that no one except the sender and intended recipients of a communication can access the contents. By scanning users' messages, extracting content and reporting it to a third party, client-side scanning breaks that fundamental premise and exposes people to major security risks.

Moreover, this type of message scanning is not effective at stopping the proliferation of child sexual exploitation and abuse or catching abusers,[42] and may actively harm children by denying them security and privacy. For example, in 'Bugs in our Pockets', a coalition of experts in cryptography, computer science and policy state that client-side scanning technology "*by its nature creates serious security and privacy risks for all society while the assistance it can provide for law enforcement is at best problematic. There are multiple ways in which client-side scanning can fail, can be evaded, and can be abused.*"[43]

There is no way to make scanning technologies work for 'good' purposes only. Revelations of hacking like the Pegasus spyware scandal demonstrate that there are already many efforts underway around the

[41] EFF: Why Adding Client-Side Scanning Breaks End-To-End Encryption
Riana Pfefferkorn: CLIENT-SIDE SCANNING AND WINNIE-THE-POOH REDUX (PLUS SOME THOUGHTS ON ZOOM)

[42] De Montoye et al: Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning

[43] H Abelson et al: Bugs in our Pockets: The Risks of Client-Side Scanning;

world by governments to invade the privacy of their own citizens as well as those of other countries. [REDACTED]

Concerns surrounding client-side scanning, and the importance of end-to-end encryption, are set out at length in the independent human rights impact assessment of Meta's end-to-end encryption plans conducted by Business for Social Responsibility, which concluded that "*the deployment of client-side scanning technologies as they exist today should not be pursued, as doing so would undermine the cryptographic integrity of end-to-end encryption and constitute a disproportionate restriction on privacy and a range of other human rights.*"[44]

[44] https://www.bsr.org/en/our-insights/blog-view/human-rights-assessment-of-metas-expansion-of-end-to-end-encryption