April 2022

Meta Response: End-to-end Encryption Human Rights Impact Assessment

Expansion of E2EE to Messenger and Instagram DMs

Meta

Table of Contents

Overview **3**
- About this HRIA 4
- Why Have an HRIA Response? 5
- Acknowledgements 5

Findings **6**

HRIA Recommendations **8**
- Overview 8
- Meta Responses to Recommendations 9
  - Product 9
  - Process 13
  - Product Policy 16
  - Public Policy 21

Overview

Billions of people around the world use messaging apps every day to keep in touch with friends and family, conduct business, send payments, access government services, and share some of their most sensitive information.

End-to-end encryption (E2EE) provides strong privacy and security guarantees to people who use these services: it ensures that only the sender and the intended recipient or recipients of a communication, and no one in between, can access, infer, or tamper with its content. This makes it impossible for even the service provider to obtain or disclose the content of someone's communications.

Such safe, secure E2EE messaging has been widely deployed in recent years. Notably, popular messaging services like iMessage, LINE, Signal, Viber, and WhatsApp all use varying forms of E2EE to secure people's communications by default.

In March 2019, Meta (then Facebook) announced plans to deploy E2EE by default across its Messenger and Instagram DM messaging services, building on its existing implementation of E2EE in WhatsApp since 2016. This effort would require a complete re-architecting of Messenger and Instagram DM services over a number of years; as of publication in April 2022, Meta plans to complete this transition sometime in 2023.

The potential benefits and positive human rights impacts of E2EE messaging have been widely acknowledged, especially with regard to privacy, freedom of expression, protection against increasingly sophisticated cybercrime threats, physical safety, and freedom from state-sponsored surveillance and espionage in an age of rising digital authoritarianism.

Stakeholders have also raised important concerns about risks and adverse impacts of E2EE messaging—notably including the impacts on child safety, ability to proactively moderate harmful content, and ability to provide law enforcement authorities with access to message content for legitimate investigations.

Yet, the impacts of E2EE go far beyond such a simplistic "privacy versus security" or "privacy versus safety" framing. E2EE messaging has far-reaching and nuanced implications for the full range of human rights. This human rights impact assessment (HRIA) is intended to provide a comprehensive overview of these impacts in the context of Meta's plans to expand E2EE across Messenger and Instagram DMs, while our response discusses how Meta aims to address these impacts and related recommendations.

About this HRIA

Alongside our March 2019 announcement, we made a promise to carry out our expansion of E2EE thoughtfully, with an eye to both providing strong privacy and security guarantees *and* safety.

Thus, in line with our corporate commitments to human rights due diligence, we commissioned this independent HRIA of risks and opportunities stemming from our plans to expand E2EE across Messenger and Instagram DMs.

The HRIA was independently led and written by Business for Social Responsibility (BSR), a non-profit human rights and sustainability consulting firm.

It's important to note that, although many of its findings may be broadly applicable to all E2EE messaging products, this HRIA does not examine E2EE messaging generally or Meta's existing products and features employing E2EE (such as WhatsApp and Messenger Secret Conversations), although many of its findings may be broadly applicable. It's also not an HRIA of Meta's plans to enable cross-app communication among its messaging services (though the HRIA makes passing reference to these nascent plans in some of its findings).

Using methodology aligned with the UN Guiding Principles on Business and Human Rights (UNGPs), BSR performed the core work of its assessment between late 2019 and late 2021, and finalized the HRIA in early 2022. The HRIA considers the impacts of Meta's E2EE expansion—both positive and negative—on a wide range of salient rights.

BSR spoke with dozens of stakeholders, including independent academics and civil society organizations with insights into the interests of rightsholders, as well as to technical experts and organizations specializing in privacy, freedom of expression, human rights defenders, addressing

violence against women, child rights and child protection, countering terrorism and violent extremism, and human trafficking prevention and enforcement. BSR's rightsholder and stakeholder consultation took the form of interviews to inform the analysis and conclusions in the HRIA, as well as peer review of the assessment.

While we typically share insights and actions from our due diligence as part of our annual human rights reporting, we are publishing this HRIA and response in full as a standalone product. We believe this HRIA and response represent a potentially groundbreaking and timely contribution to the ongoing conversation around the deployment of E2EE, and can significantly inform a human rights-based analysis of this topic. This exceptional disclosure is in line with the approach outlined in our Corporate Human Rights Policy to publish in full due diligence that we believe "meaningfully advances the human rights field."

This disclosure is also part of Meta's broader commitment to meaningful transparency about our human rights due diligence, and about our product and policy integrity work.

Why Have an HRIA Response?

This Meta response is intended as a summary of the HRIA process and findings, and as a guide to what we have done and will do to follow up.

By sharing this HRIA and our response, we're seeking to fulfill the expectations of Principle 21 of the UNGPs and the commitments made in our Corporate Human Rights Policy.

Good human rights due diligence is not just a compliance exercise. We are actively seeking to learn from this HRIA to inform our work, to mitigate potential risks in our products and policies, and to serve users around the world.

Acknowledgements

We greatly appreciate the work and insights of this HRIA, and are deeply grateful to the many experts, defenders, academics, Meta employees, and others who provided input. Doing human rights due diligence on highly technical product changes is very challenging—all the more so given the complex nature of encryption technologies, and the many pressures of related public conversation.

We are especially grateful to the human rights and technical experts who generously offered their time to the BSR team to peer review a pre-publication draft of their HRIA.

Findings

BSR's assessment underscores the opportunities presented by our expansion of encryption across our messaging apps, including enabling:

- privacy and its additional benefits of, among many others, free expression, opinion, association, movement, religion, and belief;
- physical safety, particularly for vulnerable communities; and
- journalistic integrity and freedom, and access to information.

In addition, BSR concludes that Meta's plan to expand E2EE leads to increased realization of a diverse range of other human rights, including access to remedy, and participation in government.

The HRIA also highlights important human rights risks—including those related to child sexual abuse and exploitation; the virality of hate speech and harmful mis/disinformation; malicious coordinated behavior and information operations; the sale of illicit goods; human trafficking; and terrorism and violent extremism.

With regard to these risks, the assessment concludes Meta's plan to expand E2EE does not, in and of itself, cause or contribute to harm, *as long as* Meta investigates and deploys reasonable integrity and safety mitigations. BSR's proposed mitigations, which are addressed in detail in our responses to the recommendations laid out below, are broadly consistent with the integrity approach to E2EE messaging on Messenger and DMs we previously announced and adopted in December 2021.

As part of the HRIA, BSR also conducted a thorough analysis of the potential human rights impacts of various proposals for "technical solutions" to proactively detect or monitor content on E2EE messaging platforms, generally referred to as "client-side scanning." Client-side scanning would involve leveraging a user's device to scan for the presence of certain harmful or

prohibited content, such as known child sexual abuse material (CSAM), and report that content to third parties such as the servivce provider or law enforcement.

As the HRIA highlights, technical experts and human rights stakeholders alike have raised significant concerns about such client-side scanning systems, including impacts on privacy, technical and security risks, and fears that governments could mandate they be used for surveillance and censorship in ways that restrict legitimate expression, opinion, and political participation that is clearly protected under international human rights law.

BSR concludes that deployment of client-side scanning technologies as they exist today should not be pursued, as doing so would undermine the cryptographic integrity of E2EE and constitute a disproportionate restriction on privacy and a range of other human rights. BSR further finds that other theoretical approaches to client-side scanning in the context of a messaging service that are not feasible with current technology, such as homomorphic encryption, would also pose important human rights risks that would need to be explored and adequately addressed before any implementation.

While BSR recommends that Meta continue investigating technologies such as homomorphic encryption that could potentially enable the detection of CSAM while preserving cryptographic integrity, it stresses both the speculative nature of this technology and the importance of conducting additional human rights due diligence on any possible approaches. BSR concludes that these technologies should not be pursued in the event that this additional due diligence were to suggest they would likely result in a significant restriction of freedom of expression and other rights.

Meta believes that any form of client-side scanning that exposes information about the content of a message without the consent and control of the sender or intended recipients is fundamentally incompatible with an E2EE messaging service. People who use E2EE messaging services rely on a basic promise: that only the sender and intended recipients of a message can know or infer the contents of that message.

HRIA Recommendations

Overview

The HRIA made 45 recommendations. At the time of writing, Meta had committed to implement 34 recommendations, partly implement 4 recommendations, and assess the feasibility of another 6. We will take no further action on 1 recommendation.

In line with the overall framing of this HRIA as described above, our responses to these recommendations relate to Meta's expansion of E2EE to Messenger and Instagram DMs; unless specifically indicated they are not intended to speak to WhatsApp or any other product.

It is also important to be transparent and note we may not implement all the recommendations of which we're assessing feasibility. We will provide an update in our future annual human rights reporting.

For ease of reading, we have categorized our response to HRIA recommendations to match the categories and order in which they appear in BSR's independent assessment. We are characterizing our responses as follows:

> **Implementing:** We have implemented or are implementing steps that are consistent with, or have otherwise satisfied, the recommendation.
>
> **Implementing in part:** We have implemented or are implementing steps that encompass, or have satisfied, certain of the recommended actions.
>
> **Assessing feasibility:** We are assessing the feasibility and impact of the recommendation and will provide further updates in the future.
>
> **No further action:** We will not implement the recommendation, either due to a lack of feasibility or disagreement about how to reach the desired outcome.

Meta Responses to Recommendations

1. Product

a) Provide more consistent, cohesive, accessible, and user-friendly methods for user reporting across messaging platforms.

Implementing

We agree with BSR that user reporting is a critical signal of abuse in a private messaging context. Our goal is to encourage significantly more reporting by making it more accessible.

As we recently announced, Meta has already taken a number of steps to simplify reporting in line with this recommendation, including increasing the prominence of reporting options, reducing the number of steps in our reporting flow, and allowing users to explicitly flag the most severe violations, such as those involving children. As of this writing, we've seen close to 50% year-over-year growth in reporting, and we're taking action to keep Messenger and Instagram DMs safe.

Meta is committed to continuing to iterate on this work to improve reporting features across Messenger and Instagram DMs. We will also explore ways to provide appropriate consistency in the user reporting experience across Messenger, Instagram DMs, and WhatsApp in the context of cross-app communications, while allowing for the distinct nature and unique features of these products.

b) Ensure that user interfaces (especially user reporting features) are easy to find, simple to use, and available in all the languages Meta supports.

Implementing

We want private messaging on Messenger and Instagram DMs to be simple and accessible for the more than two billion people who use one of these apps every month. As we recently announced, and as described above in our response to recommendation 1(a), we are heavily focused on this work as part of our approach to safety and integrity in E2EE messaging.

Our reporting flows are available in 61 languages as of April 2022, including all languages that our apps are fully translated into. We do not restrict the languages that people can use in their private messages, but continually re-evaluate which of the thousands of languages spoken and written around the world to formally support with translations and content moderation resources based on factors including prevalence, emerging risks, and technical feasibility.

c) To protect children from unsolicited interactions with adults (which might lead to grooming and trafficking), Meta's UX/UI Research group should conduct participatory and co-design workshops to test user reporting features with children.

Implementing

We are conducting ongoing UX/UI research with children and consulting global online safety experts and child safety organizations for insight and feedback as we continue to improve our reporting features to make them more accessible, comprehensible, and globally relevant for children.

For example, we've learned from these consultations that it's important that we increase the visibility of reporting tools for children. This is why we've launched features like safety notices on Messenger and Instagram DMs that display a pop-up educational message and suggestions to report or block in a messaging thread when we detect behavioral signals that indicate suspicious activity. We're currently testing other reporting reminders like this inside our products to encourage reporting during potentially critical moments.

We've also added the ability to select "Involves a child" in some of our relevant reporting tools, which, in addition to other factors, prioritizes the report for review and action. Our goal is to encourage significantly more reporting by making it more accessible, especially among young people. As of this writing, we've seen close to 50% year-over-year growth in reporting, and we're taking action to keep Messenger and Instagram DMs safe.

We are committed to conducting participatory user research and co-design, including with child rights stakeholders in civil society and children, to inform this work.

d) Develop documentation and measurement techniques to assess the degree to which user reporting is helping keep users safe online.

Implementing

We continuously assess the effectiveness of our policies, processes, and tools. In line with this long-standing practice, we will continue to assess the effectiveness of user reporting to inform our integrity efforts across E2EE messaging.

e) Explore and define how to verify the authenticity of user reports.

Implementing

We agree that employing robust techniques to verify the authenticity of reports is important to preventing abuse, including fake / malicious reporting, in the context of E2EE messaging. We currently employ a variety of techniques to verify such reports. One such technique is "message franking," which we have employed in Messenger Secret Conversations since it first launched in 2016 to securely authenticate messages reported to us on Messenger (as further described in a public whitepaper).

We will consider the pros and cons of various techniques for verifying user reports as we expand default E2EE across Messenger and Instagram DMs, while bearing in mind the distinct nature and unique features of our messaging services

f) Invest in processes to ensure that users who have violated platform policies cannot return.

Implementing

Meta already has strong policies addressing recidivism by users who have previously violated our Community Standards across Messenger and Instagram, and we explicitly prohibit the creation of new accounts to bypass previous restrictions or bans.

Recidivism is a highly adversarial space, with many opportunities for sophisticated and motivated actors to take steps to evade detection. We are committed to continuing to invest significant resources in countering these efforts and working to prevent recidivism through a combination of automated and human enforcement. Given that our approach to recidivism is already largely reliant on metadata, we expect that we will continue to be able to take strong action across E2EE messaging services.

g) Expand and simplify in-app support and education features for vulnerable groups, such as children or those with lower levels of digital literacy.

Implementing

Meta consistently assesses the support and education we provide to our users, including vulnerable groups. We are committed to continuing to expand contextual safety features.

For example, we recently began to provide users with in-the-moment education to suggest blocking someone when they receive messages on a linked account from someone they've already blocked on another app. In addition, we recently simplified the descriptions in our reporting flows, reduced the number of steps involved, and took steps to emphasize that reporting is private. We've also recently improved our safety notices that prompt children to take action to block, report, or ignore/restrict someone when something doesn't seem right by making sure the user cannot continue messaging without first interacting with an educational module.

h) Assess options for "friction" when contacting groups and strangers on messaging platforms in order to minimize unsolicited interactions, virality of harmful mis/disinformation and hate speech, harmful coordinated behavior, and other actions that may lead to adverse human rights impacts.

Implementing

We are committed to continuing to implement friction across Messenger and Instagram DMs to address harmful behavior.

Our machine learning technology will look across non-encrypted parts of our platforms — like account information and photos uploaded to public spaces — to help detect suspicious activity and abuse.

For example, if an adult repeatedly sets up new profiles and tries to connect with children they don't know or messages a large number of strangers, we can intervene to take action, such as preventing them from interacting with children. We can also default children into private or "friends only" accounts on Facebook and Instagram. We also educate young people with in-app advice on avoiding unwanted interactions.

i) Only implement end-to-end encryption on Messenger Kids and Instagram for Kids if it is possible to retain the same amount of parental control that is currently available.

Implementing

While we have not yet determined how and whether to implement E2EE in Messenger Kids, we are committed to maintaining the same strong parental controls in Messenger Kids as we expand E2EE, including the ability for parents to control who their children can message, and when and whether they can use the app.

As announced in September 2021, we have indefinitely paused the launch of Instagram Kids to consult further with parents, experts, and policymakers. Our approach to parental control in E2EE environments is an important part of this ongoing consultation.

j) To protect the privacy and anonymity of users, account linking should not be mandatory and users should have different options to opt-in or opt-out upon registering and using WhatsApp, Instagram DMs, and Messenger.

Implementing

Meta does not and has no plans to require users to link their existing accounts across Messenger, Instagram, and WhatsApp if they do not wish to do so. Linking of multiple existing accounts across Meta messaging services will be voluntary and require affirmative opt-in from users.

2. Process

a) Continue to invest in harm prevention strategies in end-to-end encrypted messaging, such as the use of metadata analysis and behavioral signals, redirection/behavioral nudges, user education, etc., and communicate publicly on lessons learned and the effectiveness of such methods for addressing different kinds of harm.

Implementing

In line with the approach to private messaging integrity on Messenger and Instagram DMs that we announced in December 2021, Meta is strongly committed to continued investment in addressing potential risks harms in E2EE messaging through a range of approaches, including, consistent with any applicable law, use of metadata and behavioral signals to help prevent, detect, and report harmful behavior; conscious product design choices; user education; and robust stakeholder engagement. This work will be extensively informed by and build on our longstanding use of these approaches across

our messaging and social networking services, as well as ongoing research, human rights due diligence, and rights holder engagement.

b) During the design and development of ML techniques to proactively detect harmful accounts and content in end-to-end encrypted messaging, follow "human rights by design" guidelines to ensure user privacy, fairness, transparency, interpretability, and auditability.

Implementing

As outlined in our response to recommendation 2(a) above, we are committed to responsible development of a variety of approaches to harms in E2EE messaging contexts. As we build out these efforts, we will do so in line with Meta's Corporate Human Rights Policy, our longstanding privacy work, and commitment to responsible development of AI.

c) Create a child rights strategy for private messaging services that brings together all the elements needed to address risks to child rights holistically.

Implementing in part

We agree with the importance of centering child rights within our approach to private messaging, and to integrity and safety more broadly. We continue to develop our approach using the framework of 'prevention', 'user control' and 'responding to harm'.

As part of the full text of this recommendation, BSR also suggests that Meta continue to evaluate client-side scanning as an approach to addressing CSAM. Meta believes that client-side scanning could lead to information exposed to third parties without users' consent or control, and is fundamentally incompatible with an E2EE messaging service. We do not plan to further pursue such approaches. For more details, please see our response to recommendation 2(d) and 2(e) below.

d) Continue investigating client-side scanning techniques to detect CSAM on end-to-end encrypted messaging platforms, in search of methods that can achieve child rights goals in a manner that maintains the cryptographic integrity of end-to-end encryption and is consistent with the principles of necessity, proportionality, and nondiscrimination.

No further action

As the HRIA highlights, technical experts and human rights stakeholders alike have raised significant concerns about such client-side scanning systems, including impacts on privacy, technical and security risks, and fears that governments could mandate they be used for surveillance and censorship in ways that restrict legitimate expression, opinion, and political participation that is clearly protected under international human rights law.

Meta shares these concerns. Meta believes that any form of client-side scanning that exposes information about the content of a message without the consent and control of the sender or intended recipients is fundamentally incompatible with an E2EE messaging service. This would be the case even with theoretical approaches that could maintain "cryptographic integrity" such as via a technology like homomorphic encryption—which the HRIA rightly notes is a nascent technology whose feasibility in this context is still speculative.

People who use E2EE messaging services rely on a basic premise: that only the sender and intended recipients of a message can know or infer the contents of that message. As a result, Meta does not plan to actively pursue any such client-side scanning technologies that are inconsistent with this user expectation.

e) If Meta identifies client-side scanning methods capable of detecting CSAM while maintaining the cryptographic integrity of end-to-end encryption, then this should only be implemented after a review of the potential adverse human rights impacts (for example, on privacy, freedom of expression) and a conclusion that those impacts could be adequately addressed.

Implementing

To the extent that such methods are identified by others and brought to Meta's attention, we will consider implementation only after careful human rights due diligence to identify potential adverse human rights impacts and appropriate mitigations.

f) Conduct human rights due diligence on cross-app communication.

Implementing

Meta will conduct human rights due diligence on cross-app communication, in line with

our Corporate Human Rights Policy. We will report on insights and actions from this due diligence as part of future annual human rights reporting.

3. Product Policy

a) Develop new privacy policies with enhanced consistency across all three messaging platforms, and be more transparent about user data collection, data retention, and data sharing.

Assessing feasibility

We share BSR's desire to ensure that our privacy policies are transparent and accessible. Messenger and Instagram DMs already share a unified Data Policy.

As we continue to develop our approach to cross-app messaging, we will assess the extent to which it is possible and appropriate to make the privacy policies for these distinct services more consistent. In doing so, we will take into account the unique features of each of our E2EE messaging platforms, including the relevance of discoverability, promoted content, and connections to the social graph.

b) Apply a minimum level of consistency in Community Standards across all three messaging platforms to facilitate improved user reporting.

Implementing in Part

Today, Messenger and Instagram DMs already share a unified baseline set of content policies, grounded in the Community Standards.

As we continue to develop our approach to cross-app messaging, we will assess the extent to which it is appropriate to further align content policies across our private messaging services. In doing so, we will take into account the unique features of each of our E2EE messaging platforms, including the relevance of discoverability, promoted content, and connections to the social graph.

c) Consult with the Oversight Board about (1) whether to maintain separate standards for each messaging platform or develop a single unified standard, and (2) what level of content standards are appropriate for Meta's private messaging services.

Assessing feasibility

The Oversight Board is fully independent of Meta and has full discretion over the cases and requests for policy guidance it accepts for review. Given that caveat, while we are not opposed to seeking the Oversight Board's non-binding guidance on this important question, Meta is not currently able to commit to fully implementing this recommendation.

As we continue to develop our approach to cross-app messaging, we will assess the extent to which it is appropriate and possible to consult the Oversight Board on this matter.

d) Apply the stricter standard in cases where separate content standards conflict (e.g., a message sent from WhatsApp to Messenger that violates Community Standards in the latter but not in the former).

Assessing feasibility

Please see our response to recommendation 3(b) above.

e) Develop publicly available, accessible, and understandable policy documents to disclose Meta's use of ML classifiers for detecting, flagging, and moderating accounts and content on messaging platforms.

Implementing

We invest in and share details of our artificial intelligence technologies to improve our ability to detect violating content and keep people safe. Whether it's improving an existing system or introducing a new one, these investments help us automate decisions on content so we can respond faster and reduce mistakes.

The challenges of harmful content affect the entire tech industry and society at large. That's why we open-source significant portions of our technology to make it available for others to use. We believe being open and collaborative with the AI community will spur research and development, create new ways of detecting and preventing harmful content, and help keep people safe.

In line with the full text of BSR's recommendation, in considering public disclosures about our use of classifiers, Meta will continue to balance transparency against the extent to

which public disclosure of certain details may allow abuse of our systems by malicious actors.

f) Examine whether and how ML classifiers for detecting, flagging, and moderating accounts and content on messaging platforms could result in discrimination.

Implementing

As part of our company-wide approach to responsible AI and our Corporate Human Rights Policy, we are committed to continuing to assess how our use of classifiers and other technologies may result in discrimination.

g) To avoid creating "black box" machine learning systems and missing potential blind spots in content moderation, undertake internal and external audits by reliable third-party organizations.

Implementing in part

In line with our company-wide approach to responsible AI and our Corporate Human Rights Policy, we are committed to continuously evaluating the effectiveness and impacts of the machine learning systems we employ for content moderation, including to identify and address potential "blind spots" in our approach.

Given the rapidly evolving nature of these technologies and the need to respond to a highly adversarial environment, we do not believe that third-party audits are appropriate or effective, and instead we will focus on agile internal assessments of our technology to inform our work.

h) Report the amount of problematic activity detected and accounts suspended on messaging platforms, as well as the success rates of the detection, disaggregated by relevant factors such as gender, geography, or age.

Assessing feasibility

We already publish regular reports on how we enforce our Community Standards and respond to government and rights holder requests, including across our messaging services.

We will continue to assess the extent to which we can provide additional disaggregated information on our enforcement efforts given technical, integrity, and privacy constraints.

i) Identify what new types of data governments may begin to request in end-to-end encrypted contexts, and form a perspective on when, how, and following what processes this data should be shared.

Implementing

Meta responds to government requests for data in accordance with applicable law and our terms of service, as outlined in our Guidelines for Law Enforcement Authorities and consistent with our commitments under our Corporate Human Rights Policy and as a member of the Global Network Initiative.

Each and every request we receive is carefully reviewed for legal sufficiency and we may reject or require greater specificity on requests that appear overly broad or vague, or that are inconsistent with international standards. Our policy is to produce information that is narrowly tailored to respond to each request. We do not retain data at the request of governments absent a legal obligation to do so (such as a valid data request or preservation request), or unless we intend to make a proactive report to help promote the safety and security of our users and products, consistent with the applicable Terms of Service.

As we expand our deployment of E2EE messaging across Messenger and Instagram DMs, we are committed to continuing to follow this careful approach to engagement with law enforcement.

j) Modify enforcement policies to account for the uncertainty around the extent to which behavioral signals "prove" that a user has violated Meta's content standards.

Implementing

We are committed to continuing to iterate on our enforcement policies to reflect the realities of E2EE messaging. As of April 2022, Meta offers appeals for the vast majority of violation types, and we thoroughly validate violations arising from E2EE conversations on Messenger and Instagram DMs.

To appeal a decision, people select the option to "Request Review" after we notify them that their content or account has been actioned. When a review is requested, Meta assesses the report again using a combination of human review and/or automation to determine whether or not the correct action was taken under our Community Standards. This process allows people to let us know if they think we've made a mistake, which is essential to helping us build a fair system.

We are beginning to provide appeals not just for content that we took action on, but also for content that was reported but not acted on.

k) Provide more information about how Meta's appeals process works in end-to-end encrypted platforms.

Implementing

In line with our existing transparency commitments, we are committed to providing ongoing transparency on how our content moderation processes—including appeals—work in the context of E2EE messaging on Messenger and Instagram DMs.

l) Increase the speed and capacity of reporting and appeals processes, especially for vulnerable groups.

Implementing

As we recently announced, Meta has already taken a number of steps to simplify reporting in line with this recommendation, including increasing the prominence of reporting options, reducing the number of steps in our reporting flow, and allowing users to explicitly flag the most severe violations, such as those involving children. We've done this with vulnerable communities in mind. As of this writing, we've seen close to 50% year-over-year growth in reporting, and we're taking action to keep Messenger and Instagram DMs safe.

As noted in our response to recommendation 1(b) above, our reporting flows are available in 61 languages as of April 2022, including all languages that our apps are fully translated into. We do not restrict the languages that people can use in their private messages, but continually re-evaluate which of the thousands of languages spoken and written around the world to formally support with translations and content moderation resources based on factors including prevalence, emerging risks, and technical feasibility.

m) Assess the grievance, reporting, and appeals process against the UNGPs effectiveness criteria for nonjudicial grievance mechanisms (i.e., legitimacy, accessibility, predictability, equitability, transparency, rights-compatible, source of continuous learning).

Implementing

We will perform this due diligence in line with our Corporate Human Rights Policy. Prioritization and timing will depend on the results of an ongoing Salient Risk Assessment. We will disclose further details on insights and actions in an upcoming Meta annual human rights report.

n) Integrate human rights due diligence into privacy reviews and data protection assessment procedures.

Assessing feasibility

Our Corporate Human Rights Policy outlines a strong commitment to human rights due diligence across our operations, in line with the UNGPs. This includes human rights due diligence related to privacy rights.

We will assess whether it may be appropriate to combine privacy-related human rights due diligence with our at-scale and highly detailed privacy and data protection assessment procedures.

4. Public Policy

a) Proactively advocate in favor of end-to-end encryption and against government hacking, and resist attempts by governments to prevent, ban, undermine, or interfere with end-to-end encryption, both alone and in coordination with others.

Implementing

Meta believes strongly in the importance of E2EE to protect the human rights of the people who use our messaging products. We will continue to work to publicly advocate in favor of E2EE and defend against attempts by governments to prevent, ban, undermine, or interfere with E2EE.

We will also continue to publicly advocate against the irresponsible use of commercial spyware, for example in our detailed 2021 report on the surveillance for hire industry, and

have taken action against spyware companies that have attempted to abuse our services and harm our users. We are deeply grateful for the efforts that civil-society organizations, advocates, and experts play in defending E2EE around the world.

b) Engage policymakers about conflicting regulatory requirements that unnecessarily pit privacy rights against protecting users from broader harm, such as the European Privacy Directive.

Implementing
In line with our response to recommendation 4(a), immediately above, we are committed to engaging with policymakers to advocate in support of E2EE messaging and against regulatory efforts to prevent, ban, undermine, or interfere with E2EE.

c) Participate actively, constructively, and collaboratively in dialogue with civil society organizations, academics, the technical community, governments, and other relevant stakeholders about methods to address the adverse human rights impacts arising from the deployment of end-to-end encryption.

Implementing
We are deeply grateful for the efforts civil society organizations, advocates, and experts play in advocating for and defending E2EE around the world. We are engaged in ongoing dialogue with civil society, academics, the technical community, governments, and other stakeholders as we define our approach to E2EE messaging in Messenger and Instagram DMs, and will continue this engagement,

d) Organize internal workshops and invite experts and academics who work on content-moderation techniques in an end-to-end encrypted environment to discuss the pros, cons, and feasibility of various mitigation techniques for specific issues.

Implementing
Please see our response to recommendation 4(c), immediately above. Meta will continue to engage with stakeholders and organize workshops and other events to discuss the important human rights impacts and integrity challenges associated with private messaging.

However, as noted in our response to recommendation 2(d) above, we will not support or pursue measures that weaken or undermine E2EE, such as client-side scanning or message tracing, as we believe them to be fundamentally incompatible with an E2EE messaging service.[1]

e) Continue to explore ways to provide data and other information for researchers focused on end- to-end encrypted messaging.

Implementing

We will continue to explore possible approaches to sharing data on E2EE messaging with researchers, consistent with our privacy commitments and applicable law. We will provide further updates on our progress in future human rights reporting.

f) Continue funding researchers who are capable of carrying out in-depth ethnographic research—especially in Global South countries—to understand user behavior and tactics of malicious users and vulnerable users on messaging services.

Implementing

Meta will continue to pursue in-depth user and other research, including in Global South countries, to understand use and abuse of our messaging services. We will do so through both internal research and support for external researchers, as part of our broader company-wide research efforts on integrity issues.

g) Continue funding and collaborating with civil society organizations to develop partnerships, tools, and resources that are particularly aimed at protecting users—especially vulnerable groups—from the potential adverse human rights impacts of end-to-end encrypted messaging.

[1] The full text of this recommendation cites suspicious link detection by WhatsApp as a form of "on-device scanning." This reflects an incorrect understanding of this tool. WhatsApp does not detect malicious links by matching them against a given database, but instead uses a regular expression tool to check the format of a link for unexpected URL characters (often associated with suspicious links/phishing attacks). This tool is hardcoded into the client app, and, if it detects suspicious characters in the URL, WhatsApp will neither suppress the link, nor report it out to a third party, nor have access to the message content in which the link appears.

Implementing

As outlined in our response to recommendation 4(c) and 4(d) above, and in line with our Corporate Human Rights Policy and approach to stakeholder engagement, we are committed to continuing engagement with civil society organizations around the impacts of private messaging.

h) Devote resources toward more accurately quantifying the scope of child sexual abuse material online and the corresponding harm to victims.

Implementing

We are fundamentally dedicated to supporting a better understanding of how and why people share child exploitative content, the harm it causes victims, and how to help prevent such criminal and destructive behavior.

Since 2018, we have been reporting on the amount of child nudity and exploitation content we remove. Additionally, in an effort to better understand the problem and build new interventions, in 2020, we studied our own cybertip reports to NCMEC. We then published our findings as they provided strong insights on potential preventive interventions not only on our platform but online more broadly and offline.

Finally, we helped drive the launch of the Technology Coalition's Project Protect, a large-scale, multi-pronged cross-industry effort with one multi-million dollar workstream dedicated solely to funding research to build crucial technological tools needed to more effectively prevent and work to eradicate child sexual exploitation online.

We will continue to dedicate resources to this very important work as we expand E2EE to Messenger and Instagram DMs.

i) Partner with children's rights organizations and educator groups to develop new children-specific training modules and tools tailored for the context of end-to-end encrypted messaging.

Implementing

We are currently in the process of redesigning our existing online Safety Center by working with online safety experts to help define and build resources that address end-to-end messaging.

We are also working with children's online safety organizations experts as well as educators to develop end-to-end messaging youth resources for Get Digital, our research-informed digital literacy and wellbeing program.

We're also continuing work to build tools for parents and guardians to help them get more involved in their teen's experiences on Instagram, working with international experts on bullying prevention, online safety, and digital literacy to develop resources for both guardians and teens. The first result of this work is our recently launched Family Center on Instagram.

In line with BSR's recommendation, we will continue to explore further opportunities for collaboration on child safety in consultation with our global network of experts and advisors.

j) Create issue-specific working groups within Meta's Safety Advisory Board and among its "trusted partners."

Implementing in part

Across our global network of safety advisors, which includes the members of our Safety Advisory Board, we have stood up numerous ongoing issue-specific working groups, including ones for youth-related issues, women's safety, suicide and self injury, eating disorders and human exploitation and trafficking. We also more informally work with issue-specific experts like those who have expertise in digital forensics, victim support, child development and more.

In consultation with the members of our Safety Advisor Board and our global network of trusted partners, we will continue to explore the opportunities for setting up additional issue-specific working groups focused on the impacts of E2EE private messaging.

k) Develop innovative methods to categorize reports and summarize their associated metadata for the National Center for Missing and Exploited Children (NCMEC).

Assessing feasibility

We have a deep and ongoing relationship with the National Center for Missing and Exploited Children; and work closely with them on innovating and improving their systems.

Our most recent work has included a multi-million dollar investment in the revamping of the case management system they use to categorize, track, prioritize and act on reports sent by industry—all with an eye to enabling them to more quickly prioritize the most serious and actionable reports.

We are always looking for ways to improve the information we provide in our reports and how we categorize them to assist NCMEC in its efforts to safeguard and support victims and prevent further harms from occurring.

We will work with NCMEC to explore ways we might pursue this specific recommendation as part of our ongoing collaboration, and provide updates in our future human rights reporting.

l) Continue to actively work with anti-trafficking organizations that have built relationships with survivor communities.

Implementing

We agree on the importance of work with trafficking organizations who have built relationships with survivor communities. Meta has been working with organizations like Polaris, Stop the Traffik, and A21 to inform our policies, tools, and support resources. For example, we have worked with trafficking organizations and experts to develop prevention campaigns and in-product interventions when people search for trafficking and prostitution related terms.

We will continue to expand these interventions to more locations and we will continue to prioritize this important work.

m) Proactively collaborate with, train, and inform law enforcement about how to achieve their objectives in an end-to-end encrypted world in a rights-respecting way, such as by detecting and prosecuting crimes using alternative sources of digital evidence. This collaboration should be done on a case-by-case basis, based on the rule of law context of the jurisdiction involved, and have limited objectives to prevent misuse of new capabilities or related adverse human rights impacts.

Implementing

Our dedicated Law Enforcement Outreach Team routinely engages with law enforcement

to ensure they understand our approach to lawful disclosure of data, including our commitment as a member of the Global Network Initiative to respect international human rights standards when reviewing requests. We will continue this longstanding engagement in the context of E2EE messaging on Messenger and Instagram DMs.

n) Continue working with other social media and internet companies to explore techniques to mitigate actual and potential human rights impacts of end-to-end encrypted messaging.

Implementing

Meta has strong, longstanding direct and multistakeholder integrity and security partnerships with our peer companies, on issues ranging from coordinated inauthentic behavior and cyberespionage to counter-extremism and human rights. Subject to user privacy constraints and applicable law, we will continue to leverage these partnerships to discuss and address integrity and human rights impacts of E2EE messaging.

o) Publicly communicate a strategy and action plan to address the adverse human rights impacts of end-to-end encrypted messaging, including progress toward achieving these recommendations over time.

Implementing

In line with our Corporate Human Rights Policy, we are committed to ongoing communication on insights and actions stemming from our human rights due diligence, including this HRIA. We will continue to provide updates on our approach to E2EE messaging, including these recommendations, in future annual human rights reporting and other venues.