

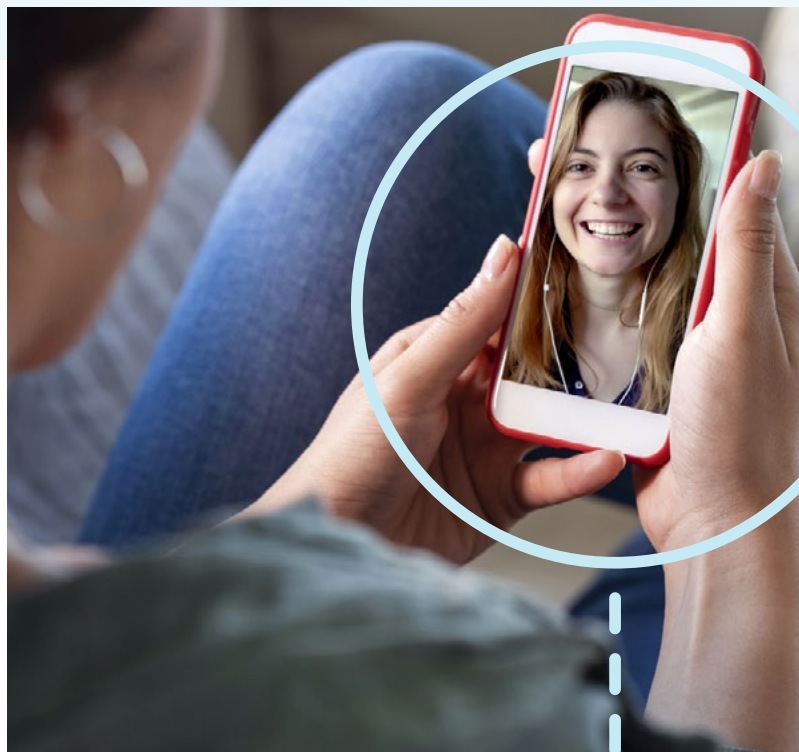


HUMAN RIGHTS IMPACT ASSESSMENT

# Meta's Expansion of End-to-End Encryption

# Contents

1. Project Overview	<b>3</b>
2. How to Read This Assessment	<b>5</b>
3. End-to-End Encryption in Context	<b>7</b>
4. Human Rights Methodology	<b>14</b>
5. Key Issues, Challenges, and Dilemmas	<b>20</b>
6. Potential Human Rights Impacts	<b>31</b>
7. Exploring the Key Human Rights Opportunities of Meta's Expansion of End-to-End Encryption	<b>42</b>
8. Exploring the Key Human Rights Risks of Meta's Expansion of End-to-End Encryption	<b>47</b>
9. Personas and Scenarios	<b>62</b>
10. Counterbalancing Competing Rights in End-to-End Encryption	<b>70</b>
11. The Human Rights Trade-offs of 'Client-Side Scanning' for Content Moderation in an End-to-End Encrypted Environment	<b>82</b>
12. Recommendations	<b>88</b>



# Project Overview

In March 2019, Mark Zuckerberg shared his view that “privacy-focused communications platforms will become even more important than today's open platforms” and that “the future of communication will increasingly shift to private, encrypted services where people can be confident what they say to each other stays secure and their messages and content won't stick around forever”<sup>1</sup>

In this post Zuckerberg described the challenges of balancing privacy and safety in the context of end-to-end encryption, and stated that Meta (formerly Facebook Inc.) will continue to discuss these challenges with experts before fully implementing end-to-end encryption across Meta’s messaging services.

Meta has three different messaging platforms—WhatsApp, Messenger, and Instagram DMs. WhatsApp is end-to-end encrypted by default; Messenger offers users the opportunity to opt-in to end-to-end encryption for each message thread; and, while optional end-to-end encrypted messaging is being publicly tested, Instagram DMs does not yet offer end-to-end encrypted messaging to all users. With over 2.8 billion users, Meta’s decision to expand end-to-end encryption



to all three messaging services (and make them capable of cross-app communication—i.e., interoperable) represents a major shift in the way the company approaches the privacy of its users and will significantly increase the use of end-to-end encrypted messaging worldwide.

In October 2019, Meta commissioned BSR to undertake a human rights impact assessment (HRIA) of extending end-to-end encryption across all Meta’s messaging services. The objectives of this HRIA are to:

- Identify and prioritize potential human rights impacts, including both risks and opportunities;
- Recommend an action plan to address the risks and maximize the opportunities;
- Inform Meta’s decisions to help ensure that end-to-end encryption is implemented in a manner consistent with human rights principles, standards, and methodologies;
- Build capacity of Meta staff and external stakeholders to understand and address the potential human rights impacts of end-to-end encryption in a messaging context.

<sup>1</sup> <https://www.facebook.com/notes/mark-zuckerberg/a-privacy-focused-vision-for-social-networking/10156700570096634/>.

It is important to note that this HRIA was undertaken in parallel with Meta’s decision-making about cross-app communication of messaging services and transition of all messaging services to end-to-end encryption. This deliberate integration of human rights into the design and decision-making phase is best practice, and is intended to help ensure that the expansion of end-to-end encryption is undertaken in a manner that avoids, prevents, and mitigates adverse human rights impacts. However, this also means that this HRIA does not include “final state” review of human rights and end-to-end encryption in Meta’s messaging services.<sup>2</sup>

This assessment was undertaken between October 2019 and September 2021. It should be noted that BSR does not make any of our own technical assertions about encryption or mitigation tactics; rather, we rely on the conclusions of technologists and cryptographers. The assessment also does not cover all the human rights implications of cross-app communication between Messenger, Instagram DMs, and WhatsApp, though elements of cross-app communication that intersect with end-to-end encryption are discussed.

## 1.1 Acknowledgments

This HRIA was conducted by Lindsey Andersen, Dunstan Allison-Hope, and Michaela Lee. BSR wishes to thank all Meta employees, rightsholders, stakeholders, and peer reviewers who participated in this assessment.

## 1.2 Disclaimer

The conclusions presented in this document represent BSR’s best professional judgment, based upon the information available and conditions existing as of the date of the review. In performing its assignment, BSR relies upon publicly available information, information provided by Meta, and information provided by third parties. Accordingly, the conclusions in this document are valid only to the extent that the information provided or available to BSR was accurate and complete, and the strength and accuracy of the conclusions may be impacted by facts, data, and context to which BSR was not privy. As such, the facts or conclusions referenced in this document should not be considered an audit, certification, or any form of qualification. This document does not constitute and cannot be relied upon as legal advice of any sort and cannot be considered an exhaustive review of legal or regulatory compliance. BSR makes no representations or warranties, express or implied, about the business or its operations. BSR maintains a policy of not acting as a representative of its membership, nor does it endorse specific policies or standards. The views expressed in this document do not reflect those of BSR member companies.

## 1.3 Suggested Citation

BSR, 2022. “Human Rights Impact Assessment: Meta’s Expansion of End-to-End Encryption.”



<sup>2</sup> Conducting a human rights impact assessment during a product design process means we assessed several product and policy decisions that may or may not ultimately be implemented. This is intentional and designed to inform Meta’s product and policy decision-making.

# How to Read This Assessment

This assessment identifies and prioritizes potential human rights impacts associated with Meta's<sup>1</sup> expansion of end-to-end encryption, considers how closely Meta is connected to these impacts, and makes recommendations for Meta to better identify, avoid, prevent, and mitigate adverse impacts. This assessment is informed by an analysis of the complex political and regulatory context in which this transition is taking place.

Individual sections of the assessment can be read separately, and key information is regularly repeated to facilitate this. However, the analysis and conclusions often build upon information in prior sections, and BSR recommends reading the assessment in its entirety to achieve a full understanding of the human rights impacts of Meta's expansion of end-to-end encryption and measures to address them.

This HRIA is organized as follows:

- **Section 3: End-to-End Encryption in Context** explains what end-to-end encryption is, why it is increasingly important to human rights, how a human rights-based approach can contribute



to the broader encryption policy debate, and what Meta's planned expansion of end-to-end encryption will look like.

- **Section 4: Human Rights Methodology** explains the methodology BSR used to conduct this HRIA, which is based on the UN Guiding Principles on Business and Human Rights (UNGPs).
- **Section 5: Key Issues, Challenges, and Dilemmas** outlines high-level observations about the key issues, challenges, and dilemmas arising from Meta's expansion of end-to-end encryption. This section provides important grounding for the rest of the assessment.
- **Section 6: Human Rights Impacts** assesses the human rights risks and opportunities associated with Meta's expansion of end-to-end encryption across the full range of international human rights. It also analyzes how Meta is connected with potential human rights impacts.

<sup>1</sup> Formerly Facebook Inc, and referring to the parent company that includes Facebook, Instagram, WhatsApp, and others.

- **Section 7: Exploring the Key Human Rights Opportunities of Meta’s Expansion of End-to-End Encryption** offers a deeper analysis of how end-to-end encrypted messaging directly enables privacy, physical safety, freedom of opinion and expression, freedom of belief and religion, and freedom of association and assembly.
- **Section 8: Exploring the Key Human Rights Risks of Meta’s Expansion of End-to-End Encryption** offers a deeper analysis of key human rights risk areas, specifically child sexual abuse and exploitation, virality of hate speech and harmful mis / disinformation, information operations, illicit goods sales, human trafficking, and terrorism, violent extremism, and hate groups.
- **Section 9: Personas and Scenarios** uses a series of hypothetical personas and scenarios to highlight how decisions Meta makes about the implementation of end-to-end encryption can disproportionately impact the rights of vulnerable groups in different contexts.
- **Section 10: Counterbalancing Competing Rights in End-to-End Encryption** outlines a methodology for addressing the numerous human rights tensions associated with end-to-end encrypted messaging. It uses this methodology to suggest human rights-based approaches to solving two of the most challenging and long-standing examples of competing rights within the broader encryption debate: the right to privacy vs. the right to security, and the right to privacy of everyone (including children) vs. the rights of children to be protected from sexual abuse and exploitation.
- **Section 11: The Human Rights Trade-offs of “Client-Side Scanning” for Content Moderation in an End-to-End Encrypted Environment** explores how any decision to detect problematic content in end-to-end encrypted messaging has significant implications for the nascent debate on content moderation in private messaging, and must be considered in the broader technical, political, and regulatory context.
- **Section 12: Recommendations** lists numerous recommendations for Meta to take appropriate actions to avoid, prevent, and mitigate the adverse human rights impacts associated with the expansion of end-to-end encryption.

# End-to-End Encryption in Context



## 3.1 Background

End-to-end encryption scrambles messages in such a way that only the sender and the recipient can decipher them. As the name implies, messages are encrypted on the device of the sender and decrypted on the device of the recipient. This means only the sender and the recipient can view or modify the messages, and even Meta, the company providing the messaging service, cannot view the contents of the messages.<sup>1</sup>

End-to-end encryption also makes it difficult for third parties to gain access to messages. For example, parties interested in seeing messages exchanged on an end-to-end encrypted platform, whether they be legitimate law enforcement actors or criminals with nefarious intentions, must go directly to a party in the conversation, have physical access to the device, or hack into the device itself via spyware or other means.<sup>2</sup> However, end-to-end encryption only protects the content of communications, not the data outside of the communications that is associated with them.<sup>3</sup>

Meta's planned expansion of end-to-end encryption to all its messaging platforms has resurfaced a

public policy debate about encryption that has been ongoing for decades.<sup>4</sup> While this debate is largely seen as pitting two opposing groups against each other in the name of two human rights in tension—privacy and security—the reality is much more nuanced.

As this HRIA will show, there are a wide range of interconnected human rights impacts arising from a move to end-to-end encryption, and it is important that Meta and other relevant actors address them in an informed, deliberate, and thoughtful manner.

## 3.2 The Growing Importance of End-to-End Encryption to Human Rights

The enhanced privacy protections enabled by end-to-end encryption are increasingly relevant for the ability of users to enjoy their human rights in practice. There are six connected reasons why end-to-end encryption should play a more central

<sup>1</sup> For more details on how end-to-end encryption works in general, see: <https://searchsecurity.techtarget.com/definition/end-to-end-encryption-E2EE>. For a technical description of WhatsApp's end-to-end encryption, see: [WhatsApp\\_Security\\_Whitepaper.pdf](https://www.whatsapp.com/legal/whatsapp-security-whitepaper)

<sup>2</sup> <https://www.nytimes.com/2019/11/19/technology/end-to-end-encryption.html>.

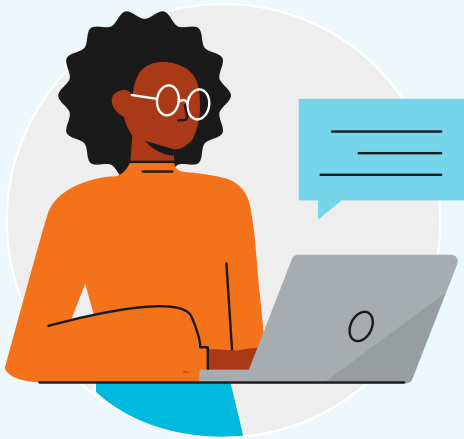
<sup>3</sup> This is a description of how end-to-end encrypted messaging works. There is also an ongoing a debate about the precise definition of end-to-end encryption, which we reference at various points throughout this assessment.

<sup>4</sup> <https://www.wired.com/story/encryption-wars-facebook-messaging/>.

role in society's strategies to protect, respect, and fulfil human rights in today's political, social, and technical context:

1. Security experts see the proliferation of end-to-end encryption as part of the natural evolution of digital security to address increasingly technically sophisticated threats.

### How End-to-End Encrypted Messaging Works



#### Sender

Jane writes message to Bob.  
Two keys are generated.  
The public key encrypts  
Jane's message.



#### Server

The encrypted message is sent to Bob through the servers of the messaging service. The messaging service cannot see the message contents.



#### Recipient

Only Bob's private key  
can unlock the message.

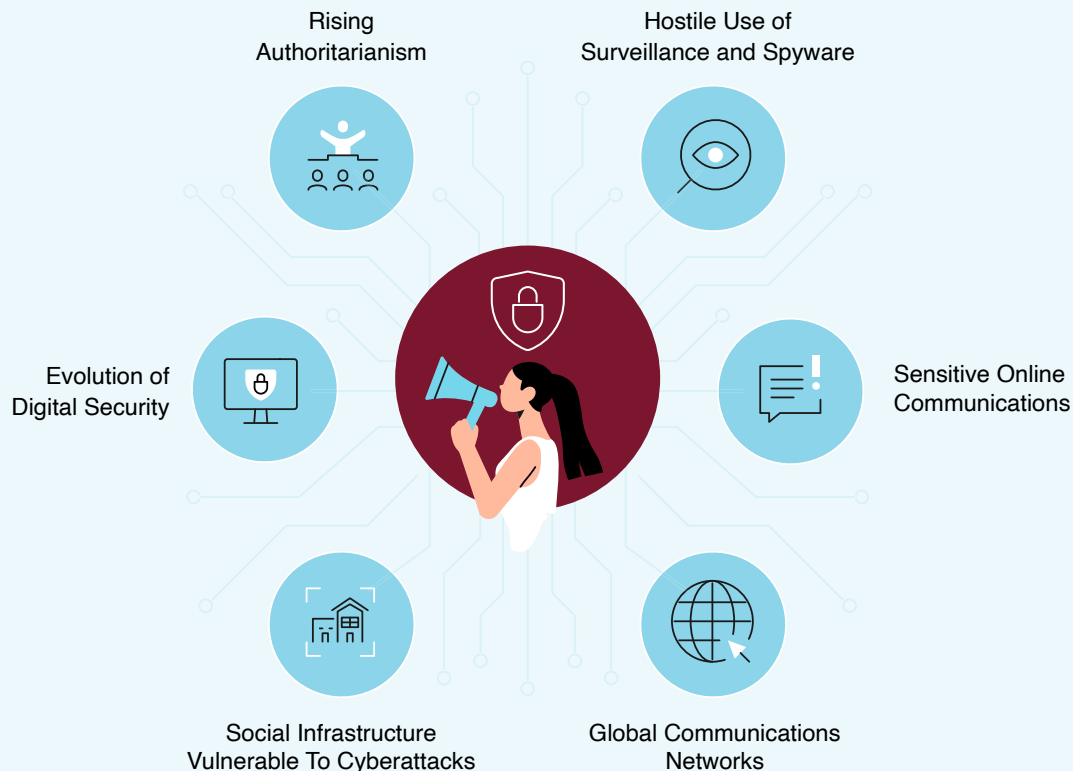


Cyberattacks are on the rise around the world as the number of threat actors, both state and non-state, who can carry out sophisticated attacks is increasing substantially. In order to defend ourselves in this context, our own security tools must evolve as well. The proliferation of end-to-end encryption is a key part of this.

**2. We are living through an age of rising authoritarianism by governments, who are placing increased restrictions on the civic space available for citizens to enjoy their rights.** The 2021 Freedom House *Freedom in the World* report found that 2020 was the 15th consecutive year of decline in global freedom, with rightsholders in the majority of countries experiencing deterioration in their political rights and civil liberties.<sup>5</sup>

**3. The strategies and tactics of authoritarianism are increasingly taking place online through surveillance, spyware, and other tactics to turn online spaces into more hostile environments that threaten human rights.** Freedom on the Net 2021 found that global internet freedom declined for the 11th consecutive year. Out of the total countries covered, more governments arrested users for nonviolent political, social, or religious speech than ever before, and authorities in at least 45 countries were suspected of obtaining sophisticated spyware or data-extraction technology from private vendors.<sup>6</sup> Freedom on the Net 2019 found that 40 countries had instituted advanced social media monitoring programs, and noted that this trend is not limited to major authoritarian powers, but is rapidly extending to smaller and poorer states too.<sup>7</sup>

### End-to-End Encryption is Essential for the Realization of Human Rights



<sup>5</sup> <https://freedomhouse.org/report/freedom-world/2020/leaderless-struggle-democracy>.

<sup>6</sup> <https://freedomhouse.org/report/freedom-net/2021/global-drive-control-big-tech>, <https://freedomhouse.org/report/freedom-net/2019/crisis-social-media>.

<sup>7</sup> <https://freedomhouse.org/report/freedom-net/2019/crisis-social-media>.

**4. We are witnessing a growth of sensitive communications taking place online, a trend that has only accelerated with COVID-19.**

Whether it is telemedicine, working remotely, or simply communicating with friends and families spread around the world, more of our private communications than ever before take place over platforms, apps, and services that rely on encryption to keep them secure.

**5. Our communications and networks and risks are increasingly global.** This means that a user in a low-risk environment—one characterized by rule of law, due process, and strong privacy protections—may communicate with a user in an environment that is anything but. Even users in high-functioning democracies can be placed at risk from the actions by governments who are not.

**6. Our social infrastructure—everything from utilities to banks and healthcare services—is increasingly vulnerable to cyberattacks by bad actors.** Catastrophic failures of digital systems would have a significant impact on our human rights, and widespread encryption (of both data in transit and data at rest) is one of the key strategies to prevent that failure from happening.

**These factors exist in a context where Meta’s family of apps has over 2.8 billion users, and is therefore a major target for bad actors.** Privacy and security while using online platforms should not only be the preserve of the technically savvy and those able to make proactive choices to opt into end-to-end encrypted services, but should be democratized and available for all.

### 3.3 How a Human Rights-Based Approach Contributes to the Encryption Debate

Meta’s planned expansion of end-to-end encryption to all its messaging platforms has resurfaced a public policy debate about encryption that has been ongoing for decades.

*There are privacy and security concerns on both sides, and there are many other human rights that are impacted by end-to-end encrypted messaging, both positively and negatively, and in ways that are interconnected.*

This debate sets two opposing groups against each other in the name of two human rights in tension—privacy and security. In this debate, a “privacy side” makes the case that end-to-end encryption provides vital protections to users in an age of mass surveillance and pushes law enforcement toward more targeted and rights-respecting intelligence and evidence gathering; meanwhile a “security side” argues that end-to-end encryption provides a safe haven for criminals, terrorists, human traffickers, and child abusers, and makes it much harder to bring these groups to justice.

The reality is much more nuanced. There are privacy and security concerns on both sides, and there are many other human rights that are impacted by end-to-end encrypted messaging, both positively and negatively, and in ways that are interconnected. It would be easy, for example, to frame the encryption debate not only as “privacy vs. security” but also as “security vs. security,” because the privacy protections of encryption also protect the bodily security of vulnerable users. End-to-end encryption can make it more challenging for law enforcement agencies to access the communications of criminals, but end-to-end encryption also makes it more challenging for criminals to access the communications of law-abiding citizens.

It is therefore important that Meta and other relevant actors address these issues in an informed, deliberate, holistic, and thoughtful manner.

The purpose of this assessment is not to focus solely on the privacy and security issues, but also to identify the myriad other positive and negative human rights impacts arising from Meta's expansion of end-to-end encryption. There are four main additional elements to keep in mind:

**First**, many human rights are potentially impacted by an expansion of end-to-end encryption. In addition to privacy and security, this assessment will describe the impact the expansion of end-to-end encryption will have on rights such as nondiscrimination, freedom of association, freedom of movement, freedom of expression, access to science and its benefits, and participation in government. We will also consider how adverse impacts should be addressed, and who has the primary responsibility to address them.

**Second**, these human rights impacts are interconnected and interrelated, which makes determining whether the impacts of end-to-end encrypted messaging are net positive or net negative an inherently flawed exercise. By providing a methodology to consider how potentially competing outcomes can be achieved at the same time, a rights-based approach provides a useful framework for stepping out of the binary “privacy vs. security” framing and more fully examining the nuances, tensions, and choices of the encryption debate.

**Third**, when considering the human rights impacts of end-to-end encrypted messaging, it is important to provide special consideration to identifying and addressing the specific needs of vulnerable groups who face heightened risks, or different risks, compared to others, and are less likely to have their needs represented in decision-making processes.

**Fourth**, it is noteworthy that opportunities arising from deploying end-to-end encryption across all

of Meta's messaging services are directly enabled by the increased privacy protections end-to-end encryption provides. By contrast, the risks arising from deploying end-to-end encryption tend to be associated with the actions of bad actors disregarding terms of service, violating the law, and adversely impacting the rights of others.

In other words, the human rights opportunities of end-to-end encrypted messaging are often first-order effects, whereas the potential human rights harms are often second- or third-order effects. This does not mean that Meta should not address harms it is not closely connected to—the UNGPs are clear that companies should address all adverse human rights impacts with which they are connected. It does, however, have implications for the leverage Meta has and the types of mitigations available.

To BSR's knowledge, this is the first-ever human rights assessment of end-to-end encrypted messaging, and we hope that the insights revealed through this assessment inform Meta's decision-making, assist other companies as they too move to adopt end-to-end encryption, and move the encryption debate in a constructive and rights-respecting direction.

### 3.4 Meta's Shift to End-to-End Encryption

Meta has three different messaging platforms—Messenger, WhatsApp, and Instagram DMs. At the time of writing,<sup>8</sup> Messenger and Instagram DMs can cross-app communicate, meaning, for example, that a user on Messenger can send a message to a user on Instagram DMs. However, WhatsApp cannot yet communicate with the other two platforms.

Of the three messaging platforms, only WhatsApp is end-to-end encrypted by default. This means that every message exchanged on WhatsApp is end-to-end encrypted automatically, including group messages, and users cannot opt out of or deactivate encryption. All content shared on WhatsApp is encrypted end-to-end, including

<sup>8</sup> September 2021.

text messages, photos, videos, voice messages, documents, status updates, and calls.<sup>9</sup>

Messenger is not currently end-to-end encrypted by default, but it does offer users the ability to encrypt their messages and video calls through the “secret conversations” feature, which is only available on the Messenger app. This feature is opt-in, and users must actively choose to encrypt each new message thread. Secret conversations encrypts text, audio and video calls, pictures, videos, voice recordings, and GIFs. At the time of writing it does not yet support payments or group messages.<sup>10</sup>

At the time of writing, Instagram DMs does not offer any end-to-end encrypted messaging capabilities, though testing with a small group of users has begun.

Although Meta is able to view the content in unencrypted messages on Messenger and Instagram DMs, Meta does not actively monitor messages. Unlike content shared to the public Facebook platform and Instagram, Meta does not proactively scan messages for content that violates the Facebook Community Standards because it views these direct messaging contexts as being private conversations.

However, there are exceptions to this for content deemed particularly harmful. For example, Meta proactively scans all images and videos to identify and remove known child sexual abuse material (CSAM), including both videos and images, using a hashing system, including as part of a legally mandated program run by the National Center for Missing and Exploited Children (NCMEC). This effort includes many other technology companies, child safety organizations, and law enforcement agencies. Meta also proactively scans for nonconsensual intimate imagery (NCII or image-based sexual assault) and particularly egregious terrorist content. Additionally, Meta does access messages in order to comply with legal requests for information

from a government or law enforcement agency in accordance with its law enforcement guidelines.<sup>11</sup>

Rather than proactively scanning for other problematic content in unencrypted messages, Meta relies on users to take action either by blocking the sender of an unwelcome message or by reporting conversations. In Messenger and Instagram DMs, users can report messages they feel violate Community Standards, even if the messages are end-to-end encrypted. If a user reports an end-to-end encrypted conversation, recent messages from that conversation are decrypted and sent to Meta for review.<sup>12</sup> WhatsApp allows users to report other users or groups, as well as specific messages. When a user submits a report, WhatsApp will receive decrypted recent messages from that person or group that were sent to the user who reported them, as well as information about those interactions.<sup>13</sup>

In shifting to end-to-end encryption for all its messaging platforms, Meta will be extending encryption from all of WhatsApp and part of Messenger to encompass all of WhatsApp, Messenger, and Instagram DMs. Whereas previously Meta could access the content of all messages on Instagram DMs and unencrypted messages on Messenger, it will no longer be able to access any message content. This extends the protections of end-to-end encryption to all users, but also means that Meta will no longer be able to provide law enforcement agencies with message content unless it has been reported by users, even when presented with a lawful judicial order that meets Meta’s law enforcement guidelines and US or Irish legal requirements.<sup>14</sup> This shift also has implications for Meta’s ability to continue proactively reporting CSAM to relevant agencies.

9 <https://faq.whatsapp.com/en/android/28030015/>.

10 <https://www.facebook.com/help/messenger-app/1084673321594605>.

11 <https://www.facebook.com/safety/groups/law/guidelines/>.

12 <https://www.facebook.com/help/messenger-app/1084673321594605>.

13 <https://faq.whatsapp.com/21197244/>.

14 Meta’s international headquarters is based in Ireland.

### 3.5 Limitations

This HRIA has been conducted while Meta has been planning for and making decisions about its end-to-end encryption rollout. Because Meta is transitioning from offering end-to-end encryption on some of its messaging platforms to all of its messaging platforms, rather than from no end-to-end encryption to full end-to-end encryption, the potential human rights impacts in relation to the status quo can be difficult to assess. This sometimes results in muddling the impacts of the specific product decision to extend encryption with the impacts of end-to-end encrypted messaging in the abstract.

While BSR has sought to provide guidance on various product and policy decisions involved in the process, some implications or decisions that affect the potential human rights impact may not have been anticipated. Additionally, because Meta's expansion of end-to-end encryption is not yet complete, this is not a final state assessment. Further assessments of potential product and policy decisions not discussed in this document may be necessary as the process continues to unfold.

#### Areas for Future Assessment

In addition to end-to-end encryption, Meta also plans to make its messaging platforms capable of cross-app communication, enabling users on one platform to message users on another. This has important implications because each platform operates in a different context with different sets of user expectations. Just as with end-to-end encryption, there are several important decisions Meta will have to make about how it implements cross-app communication when that point is reached. Meta has announced that account linking will be optional,<sup>15</sup> and that users on one platform will be able to control whether they can be contacted by others on the other platforms,<sup>16</sup> but will users be able to control what kind of information about them is searchable by others?

Will the Community Standards, currently applied to Messenger and Instagram DMs, apply to WhatsApp too? If not, what about messages sent between platforms—for example, from Messenger to WhatsApp? While this is not a HRIA of messaging cross-app communication, many decisions made about cross-app communication will affect the human rights impacts of end-to-end encrypted messaging, and are therefore considered where relevant in this assessment.

There are also a variety of potential technical mitigation measures to address the human rights risks associated with end-to-end encryption that are not technically feasible today or have not yet been proposed. One example of this is homomorphic encryption, an approach to scanning content in its encrypted form that has been proposed as a mechanism to enable Meta to detect certain kinds of harmful and illegal content in end-to-end encrypted messaging. We discuss homomorphic encryption and other proposed content scanning solutions throughout this assessment, and particularly in Sections 10 and 11. However, because it is a nascent solution that is not yet technically feasible at scale in an end-to-end encrypted messaging context, our analysis and conclusions about it are inherently speculative. Further human rights assessment should be conducted if and when such solutions are technically feasible and can be more concretely explored.

<sup>15</sup> <https://thenextweb.com/news/whatsapp-opt-in-messenger-facebook-integration>.

<sup>16</sup> <https://www.facebook.com/help/messenger-app/2258699540867663>.

# Human Rights Methodology

This HRIA was undertaken using methodologies based on the UN Guiding Principles on Business and Human Rights (UNGPs), including a consideration of the various human rights principles, standards, and methodologies on which the UNGPs were built. The HRIA helps fulfil Meta’s Corporate Human Rights Policy, which commits Meta to respecting human rights and carrying out human rights due diligence as laid out by the UNGPs<sup>1</sup>

## 4.1 Significance

With over 2.8 billion users, Meta’s various platforms and products provide a wide range of services to people who could be anywhere in the world and who may speak any language. While other end-to-end encrypted messaging services exist, Meta’s expansion of end-to-end encryption represents a major shift in the way the company approaches the privacy of its users.

Meta’s decisions will set a precedent for privacy norms across the industry. However, BSR has sought to look not just at the impacts on privacy in isolation, but also at the full range of other rights



that could be impacted. By applying various human rights methodologies and principles— including assessing impacts on all human rights, prioritizing based on severity for rightsholders, paying special attention to vulnerable groups, considering connectivity between rights, and counterbalancing potentially competing rights—we hope to contribute to the evolving field of product-level human rights due diligence and application of the UNGPs at social media companies.

## 4.2 Identifying Human Rights Impacts

The UNGPs set out the “state duty to protect human rights” and the “corporate responsibility to respect human rights.” The UNGPs expect companies to respect all human rights, as it is understood that a business decision can potentially impact any of them. Furthermore, all human rights are indivisible, interdependent, and interrelated—the improvement of one right facilitates advancement of the others; the deprivation of one right adversely affects others.

<sup>1</sup> See <https://about.fb.com/wp-content/uploads/2021/04/Facebooks-Corporate-Human-Rights-Policy.pdf>.

The expansion of end-to-end encryption will bring both negative and positive impacts that should be considered in a nuanced way. Principle 11 of the UNGPs states that businesses “should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved,” but also that companies “may undertake other commitments or activities to support and promote human rights, which may contribute to the enjoyment of rights.” However, Principle 11 also makes clear that positive impacts do not “offset a failure to respect human rights.”

For these reasons, it is important to note that (1) when we list positive impacts in this assessment, they are not being balanced or offset against adverse impacts, and (2) many of the positive impacts themselves address actual adverse impacts associated with the absence of end-to-end encryption.

In this HRIA, BSR identifies potential human rights impacts using the universe of rights codified in the following international human rights instruments:<sup>2</sup>

- The Universal Declaration of Human Rights (UDHR)
- The International Covenant on Civil and Political Rights (ICCPR)
- The International Covenant on Economic, Social and Cultural Rights (ICESCR)
- The International Convention on the Elimination of All Forms of Racial Discrimination (ICERD)
- The Convention on the Elimination of All Forms of Discrimination against Women (CEDAW)
- Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT)
- International Convention on the Protection of the Rights of All Migrant Workers and Members of Their Families (ICMW)
- Convention on the Rights of Persons with Disabilities (CRPD)
- The eight International Labour Organization (ILO) Core Conventions
- The Convention on the Rights of the Child (CRC)

### 4.3 Rightsholder and Stakeholder Consultation

An HRIA should involve meaningful engagement with rightsholders—people whose human rights may be impacted by the company—with particular attention to human rights impacts on individuals from groups or populations that may be at heightened risk of vulnerability or marginalization. Where direct engagement with these rightsholders is not possible, the UNGPs suggest that companies should use reasonable alternatives, such as engaging with independent expert resources, human rights defenders, and other representatives from civil society.

The expansion of end-to-end encryption will impact over 2.8 billion Meta users, as well as other rightsholders who do not use Meta platforms but whose rights might be impacted by those using Meta’s messaging services. Because it is impossible to engage directly with a representative cross section of individual rightsholders impacted by Meta around the world, BSR sought to understand the impacts of Meta’s expansion of end-to-end encryption by consulting independent academics and civil society organizations with insights into the interests of rightsholders, including technical experts and organizations specializing in privacy, freedom of expression, human rights defenders, addressing violence against women, child rights and child protection, counterterrorism and violent extremism, and anti-human trafficking. This rightsholder and stakeholder consultation took the form of interviews to inform the analysis and conclusions in this assessment, as well as peer review of the assessment and executive summary.

<sup>2</sup> For an explanation of international human rights law see: <https://www.ohchr.org/en/professionalinterest/pages/internationallaw.aspx>, and for background on the core international human rights instruments, see: <https://www.ohchr.org/EN/ProfessionalInterest/Pages/CoreInstruments.aspx>.

To ensure correct understanding of the technical aspects of end-to-end encrypted messaging and relevant product features and mitigations, BSR engaged with both independent technical experts and relevant product teams within Meta to inform and review the assessment. BSR does not make any of our own technical assertions; rather, we rely on the conclusions of the broader community of technologists and cryptographers and describe relevant disagreements among technical experts where it exists.

BSR also engages with a diverse range of rightsholders and stakeholders when undertaking human rights due diligence for companies across all industries. BSR supplemented the stakeholder inputs described above with our own insights into the human rights concerns of rightsholders and stakeholders gathered in a variety of contexts, including previous HRIAs undertaken for Meta.

Future human rights assessments related to end-to-end encrypted messaging, particularly certain features or mitigation measures, may benefit from direct rightsholder engagement in addition to engaging experts and those with insights into the interests of rightsholders.

#### 4.4 Prioritizing Human Rights Impacts

Principle 24 of the UNGPs acknowledges that while companies should address all their adverse human rights impacts, it is not always possible for companies to address them simultaneously, and companies should “begin with those human rights impacts that would be most severe.” This HRIA draws upon the human rights concepts of severity and vulnerable groups to prioritize the adverse impacts and the actions needed to address them.

##### Severity: Scope, Scale, and Remediability

There are three main criteria for assessing severity:

- **Scale**—The seriousness of the harm for the victim.
- **Remediability**—The extent to which remedy will restore the victim to the same or equivalent position before the harm.

In the context of social media platforms, where billions of users are also rightsholders, it is challenging to conclusively determine the scope, scale, or remediability of potential impacts. This is compounded by the fact that nonusers can also be rightsholders—for example, individuals whose rights are adversely impacted by those using Meta’s private messaging services.

Scope may be inferred by reviewing current volume trends and estimating the number of rightsholders who may be affected. However, given the sheer number of Meta users, the scope of the harm is almost always likely to be very large.

Scale and remediability are much more difficult to extrapolate in the context of social media because the seriousness of the human rights harm suffered by a rightsholder varies significantly according to the context in which it occurs. Thus, BSR has considered these criteria at a high level and with an understanding that they may vary from case to case, based on the product, feature, and policy decisions made by Meta.

BSR also typically considers the **likelihood** of the potential impact on rightsholders occurring in the next five years. However, the high-level nature of this report means that discussing the likelihood of a potential impact provides limited value to assessing human rights risks. We note that (1) there is certainty that bad actors will exploit end-to-end encrypted messaging, (2) problematic content will certainly be shared, (3) nearly all of the individual human rights risks we explore are highly likely to occur, and (4) many already exist on Meta’s current end-to-end encrypted messaging platforms, namely WhatsApp and Messenger Secret Conversations.

- **Scope**—The number of people affected by the harm.



## Vulnerable Groups

As established in the UNGPs, companies should pay “particular attention to the rights and needs of, as well as the challenges faced by, individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized.” Vulnerable groups generally face heightened risks, or different risks, compared to others, and are less likely to have their needs represented in decision-making processes. In the context of end-to-end encryption, these groups may be disproportionately impacted by the adverse human rights impacts of end-to-end encrypted messaging, but may also stand to gain the most from the human rights benefits.

Typically, vulnerable groups include, but are not limited to, human rights defenders, journalists, political dissidents, environmental and community activists, women, children, members of ethnic and religious minorities, indigenous groups, the elderly, members of the LGBTQIA+ community, and those who are illiterate or digitally illiterate. However, vulnerability depends on context, and someone who may be powerful in one context may be vulnerable in another. For this reason, BSR’s human rights methodologies are based on four dimensions of vulnerability:

- **Formal Discrimination**—Laws or policies that favor one group over another.
- **Societal Discrimination**—Cultural or social practices that marginalize some and favor others.
- **Practical Discrimination**—Marginalization due to life circumstances, such as poverty.
- **Hidden Groups**—People who might need to remain hidden and consequently may not speak up for their rights, such as undocumented migrants and sexual assault victims.

Additionally, vulnerability is heavily impacted by geographic context. In countries with a history of widespread human rights violations and/or conflict, vulnerable groups are especially at risk. There also

may be vulnerable groups outside of the typical categories that are vulnerable specifically in that geographic context.

## 4.5 Determining Appropriate Action

BSR’s HRIA methodology considers the appropriate action for a company to address adverse human rights impacts using factors contained in Principle 19 of the UNGPs.

First, we consider **attribution**, which assesses how closely connected the company would be to the human rights impact, where connection is determined using the following factors:

- **“Caused”** the impact—The company should take the necessary steps to cease or prevent the impact.
- **“Contributed”** to the impact—The company should take the necessary steps to cease or prevent its contribution and use its leverage to mitigate any remaining impact to the greatest extent possible.
- **“Directly linked”** to the impact through its products, services, or operations arising from its business relationships, including with users—The company should determine action based on factors such as the extent of leverage over the entity concerned and the severity of the abuse.

Second, we consider **leverage**, which assesses the ability of the company to affect change in the wrongful practices of an entity that causes harm, and ways to increase leverage, such as by collaborating with other actors.

## Counterbalancing Competing Rights

All human rights are indivisible, interdependent, and interrelated. The improvement of one right can facilitate advancement of others; the deprivation of one right can adversely affect others. For example, privacy is a necessary condition for the realization, promotion, and protection of many other human

rights, such as the rights to freedom of expression, freedom of assembly and association, freedom of movement, and freedom of belief and religion.

However, human rights can be in tension with one another for legitimate reasons, and rights-based methods can be deployed to define a path forward when two competing rights cannot both be achieved in their entirety. Rather than “offsetting” one right against another, it is important to pursue the fullest possible expression of both rights and identify how potential harms can be addressed.

In this assessment we used a methodology known as “counterbalancing” to identify ways to secure the fullest possible expression of rights without unduly limiting others by applying established international human rights principles such as legitimacy, necessity, proportionality, and nondiscrimination. This methodology is consistent with the notion that most human rights are not absolute, and can be limited in certain legitimate circumstances.

The encryption debate has endured for so long in part because it involves many instances of competing rights—on the one hand, end-to-end encrypted messaging protects privacy, enhances security, and enables freedom of opinion, expression, movement, association, religion, and belief; on the other hand, end-to-end encryption can hinder some efforts to protect child rights, liberty, safety, and personal security. Defining how to balance these competing rights is challenging,

*Human rights can come into conflict with one another for legitimate reasons, and rights-based methods can be deployed to define a path forward when two conflicting rights cannot both be achieved in their entirety.*

particularly because there is no definitive hierarchy of human rights—none can be considered more important than others.

Counterbalancing is not a part of the UNGPs, which do not focus on how companies should address instances of competing rights. Because competing rights are the source of so many tensions related to end-to-end encryption, we turned to international human rights law and developed a counterbalancing methodology inspired by similar exercises conducted by human rights courts. Our approach to counterbalancing in this HRIA is merely illustrative, and is shaped by the following established international human rights principles:

- **Legitimacy**—Restrictions to a right must pursue an objectively legitimate purpose and address a precise threat.



- **Necessity and proportionality**—Only restricting a right when the same goal cannot be achieved by other means, and using restrictions that are the least intrusive to achieve the legitimate purpose.
- **Nondiscrimination**—Restrictions to a right must be implemented in a nondiscriminatory manner.
- **Reverting to principle**—Focusing on the underlying principle of the right being restricted and identifying ways to uphold the core principle, even if not the exact right.

BSR undertook this human rights review from October 2019 to September 2021 following four main phases described in the table below. For reasons of timing and logistics, key elements of these project segments were undertaken concurrently, rather than sequentially.

BSR’s research and analysis took place while Meta’s decisions about how to expand end-to-end encryption in practice were still evolving, and this means that the assessment does not include a “final state” review of human rights and end-to-end encryption in Meta’s messaging services.

## 4.6 Project Phases

Phase	Activities
<p><b>IMMERSION</b></p> <p>Increase familiarity with end-to-end encryption and relevant human rights</p>	<ul style="list-style-type: none"> <li>• Review of relevant public and non-public Meta literature.</li> <li>• External literature review (e.g., regulatory context, encryption standards and protocols, civil society perspectives, academic papers).</li> <li>• Interviews with Meta engineering, technical, and policy staff to better understand the encryption transition and evolution of product road maps and features.</li> </ul>
<p><b>MAPPING AND PRIORITIZATION</b></p> <p>Engage external stakeholders to identify and prioritize potential adverse and positive human rights impacts arising from the expansion of end-to-end encrypted services</p>	<ul style="list-style-type: none"> <li>• Interviews with external stakeholders representing a diverse range of interests, geographies, and perspectives, including privacy and freedom of expression advocates; human rights defenders; child rights organizations; counterterrorism experts.</li> <li>• Use BSR’s human rights assessment methodology to map and prioritize human rights.</li> </ul>
<p><b>INITIAL DRAFT</b></p> <p>Create an initial HRIA report, including feedback from Meta and external experts</p>	<ul style="list-style-type: none"> <li>• Create first draft of BSR report, including recommendations for how to address human rights impacts.</li> <li>• Review and comment by Meta.</li> <li>• Peer review and comment by external experts.</li> <li>• Create revised BSR draft.</li> </ul>
<p><b>FINALIZE</b></p> <p>Finalize the report and communicate</p>	<ul style="list-style-type: none"> <li>• Create final draft.</li> <li>• Learning workshops and presentations to relevant Meta staff.</li> </ul>

# Key Issues, Challenges, and Dilemmas

There are several issues, challenges, and dilemmas about Meta’s expansion of end-to-end encryption that influence the conclusions and recommendations of this assessment. The following observations influence the remainder of this report.



## 5.1 Encryption Context

- **There are many human rights tensions at stake in the encryption debate.** There are far more human rights impacted by end-to-end encrypted messaging than are reflected in the dominant “privacy vs. security” framing, including rights as diverse as nondiscrimination, freedom of association, freedom of movement, freedom of thought and opinion, bodily integrity, the right to participate in government, and the right to share in scientific advancement and its benefits. These interests are interrelated and interdependent, and there are no easy or clear answers to resolving the tensions between rights. There is no hierarchy between qualified human rights, and thus no right can be privileged over another. From a human rights perspective, this means neither side of the encryption debate is “right.”
- **Meta’s expansion of end-to-end encrypted messaging will directly result in the increased realization of a range of human rights, and will address many human rights risks associated with the absence of ubiquitous end-to-end encryption on messaging platforms today.** The provision of end-to-end encrypted messaging by Meta directly enables the right to privacy, which in turn enables other rights such as freedom of expression, association, opinion, religion, and movement, and bodily security. By contrast, the human rights harms associated with end-to-end encrypted messaging are largely caused by individuals abusing messaging platforms in ways that harm the rights of others—often violating the service terms that they have agreed to. However, this does not mean that Meta should not address these harms; rather, Meta’s relationship to these harms can help identify the types of leverage Meta has available to address them.

*The provision of end-to-end encrypted messaging by Meta directly enables the right to privacy, which in turn enables other rights such as freedom of expression, association, opinion, religion, and movement, and bodily security.*

- **As the parent company for some of the dominant messaging apps, Meta is a major target for bad actors and governments trying to exploit or take action against end-to-end encryption.** Bad actors and opportunists use messaging apps to cause human rights harms at a large scale. The size of Meta’s user base makes it a target for a wide range of actors interested in influencing public sentiment, grooming and sexual abuse and exploitation of children, exchanging illegal goods and content, or sharing other content that violates Meta’s product policies. This also makes it a focal point for policymakers concerned about end-to-end encryption.
- **If Meta decided not to implement end-to-end encryption, the most sophisticated bad actors would likely choose other end-to-end encrypted communications platforms.** Sophisticated tech use is increasingly part of criminal tradecraft, and the percentage of criminals without the knowledge and skills to use end-to-end encryption will continue to decrease over time. For this reason, if Meta chose not to provide end-to-end encryption, this choice would likely not improve the company’s ability to help law enforcement identify the most sophisticated and motivated bad actors, who can choose to use other end-to-end encrypted messaging products.
- **Global perceptions of privacy and safety are evolving, particularly as the COVID-19 pandemic accelerates the shift toward increased online interactions.** In the context of rising government and corporate surveillance, users in many contexts are increasingly aware of and concerned about the privacy of their information vis-a-vis both private companies and governments. Widespread concerns about contact tracing apps and other tech-based COVID-19 responses have contributed to this. At the same time, companies globally are facing increased expectations to address user abuse of their messaging platforms in ways that lead to real world physical harm.
- **User expectations, and therefore informed consent, varies based on the product.** Facebook and Instagram started as open social network platforms with established Community Standards and guidelines, and Messenger and Instagram DMs were later added as messaging features to these already established open platforms. By contrast, the nature and purpose of WhatsApp has always been private peer-to-peer messaging of a type similar to SMS-based telecommunications services. This makes a notable difference when it comes to user expectations and informed consent across a range of topics (such as privacy, content standards, purpose), as well as aspects such as product design choices and Meta’s capacity to handle misuse and abuse of the platforms.
- **Content removal is just one way of addressing harms. Prevention methods are feasible in an end-to-end encrypted environment, and are essential for achieving better human rights outcomes over time.** The public policy debate about end-to-end encryption often focuses heavily or exclusively on the importance of detecting and removing problematic, often illegal content from platforms, whether that be CSAM or terrorist content. Content removal is important for

*Content removal is just one way of addressing harms. Prevention methods are feasible in an end-to-end encrypted environment, and are essential for achieving better human rights outcomes over time.*

many reasons. For example, every time CSAM is shared it is a repetition of harm to the victim, and therefore detecting, blocking, and removing it is key to addressing that harm. However, content removal is also a reaction to harm that has already occurred (such as the sexual abuse of a child), and does not do enough to prevent that harm from occurring in the first place. Content detection and removal in an end-to-end encrypted environment also pose numerous human rights risks, as well as practical challenges. However, there are many things Meta can do to prevent

harm from occurring in end-to-end encrypted messaging through the use of behavioral signals, public platform information, user reports, and metadata to identify and interrupt problematic behavior before it occurs.

- **There is no consensus on the degree of content moderation companies should undertake on messaging services.** This extends beyond social media companies to include telecommunications companies that provide SMS services, device companies that provide on-device messaging and standalone messaging apps, companies that provide e-mail clients, and apps that tie into social networks or other communications platforms. The content shared can vary widely, including text, still images, and video, as well as live audio calls and video calls. While there is increasing consensus about the boundaries of content moderation of content posted to open platforms such as Facebook and Instagram, this has not yet extended to messaging contexts. This dilemma will be especially relevant for Meta given the different content policies that currently apply across its three messaging platforms.



## 5.2 Impact Factors

- **The human rights impacts of expanding end-to-end encryption will vary according to geographic context.** Rightsholders who live in countries that have poor human rights records, lack the rule of law, or are in a state of conflict face increased levels of human rights risk, and in these contexts both the risks and opportunities of end-to-end encryption are likely to be amplified. Other geographic factors include languages, information ecosystems, and type of devices available. It is also important to note that even in countries with better civil and political rights records, end-to-end encryption protects people from excessive government surveillance enabled by requesting access to user information.
- **The mix of human rights risks and opportunities arising from end-to-end encrypted messaging is also highly dependent on geographic context.** In countries with extensive surveillance regimes, the main impact of end-to-end encrypted messaging may be to provide users with more options for secure communication, thereby increasing respect for rights such as privacy, freedom of expression, assembly, and association. By contrast, in countries without extensive surveillance regimes but with significant ethnic or community conflict, the main impact of end-to-end encrypted messaging may be to increase the spread of hate speech and incitement to violence in harder-to-detect formats. Meta's messaging products are also used differently in different contexts. For example, Messenger is more popular in some countries and regions than others, and certain types of problematic content such as CSAM are more frequently detected in some regions.
- **Vulnerable groups are disproportionately affected by both the negative and positive human rights impacts.** The rights of individuals from vulnerable groups and marginalized populations are disproportionately impacted

*Approaches to end-to-end encrypted messaging need to be designed with a wide range of users in mind, not simply those in affluent markets or circumstances, and not only those passing minimum age requirements.*

by the actions of others, such as authoritarian governments or other bad actors with nefarious intent. For this reason, a human rights-based approach to end-to-end encrypted messaging needs to pay special attention to the circumstances of vulnerable groups—such as the use of lower-quality devices (e.g., power, processing capability), lower levels of digital literacy, and the use of languages that Meta does not support—and it should also reflect the reality that significant numbers of children below the age of 13 use private messaging services, despite minimum age requirements. Approaches to end-to-end encrypted messaging need to be designed with a wide range of users in mind, not simply those in affluent markets or circumstances, and not only those passing minimum age requirements.

- **Meta has varying levels of resources allocated to research, investigate, and mitigate risks.** Meta's messaging services are available in almost every country in the world, but multiple factors affect the company's ability to address human rights impacts in certain regions or for certain groups. For example, some regions may have more in-country personnel, language and translation services, moderation capacity, or technical interventions than others.



### 5.3 Product Policy Factors

- **There is a debate about the definition of end-to-end encryption, and therefore what constitutes breaking or weakening of end-to-end encryption.** One side is based on a narrow definition focused on cryptographic integrity and the technical process involved in end-to-end encryption, while the other side is based on the principles behind end-to-end encryption, specifically that only the sender and intended recipients should know or infer the content of a message.

The former definition is more traditional, but has sometimes been used by those seeking “work-arounds” to detect content, while the latter is newer, but more aligned with the views of experts in the privacy and security community today.<sup>1</sup> This difference has resulted in opposing views about the validity of various proposed methods of client-side scanning—particularly those involving homomorphic encryption (which allows the processing of data while it is encrypted—see Sections 8 and 10)—that could allow the detection of harmful content such as CSAM.

Because homomorphic encryption maintains the cryptographic integrity of the underlying message content, some who utilize the narrow definition of end-to-end encryption do not believe that using it for content detection would weaken or break end-to-end encryption. However, those who utilize a broader definition argue that end-to-end encryption means that all information about the content of a message is known only to the sender and intended recipients, and therefore any system seeking to detect content and reveal information about it to a third party, even methods that maintain the cryptographic integrity of the underlying message, would “break” end-to-end encryption.<sup>2</sup>

Since this is an ongoing debate within the technical community, in this assessment BSR does not reach a point of view about whether a narrow definition of end-to-end encryption (focused on cryptographic integrity) or broad definition (focused on who knows about the content of a message) should be adopted; rather, we consider the human rights impacts of all options.

<sup>1</sup> See, for example: <https://cdt.org/insights/report-outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems/> and <https://datatracker.ietf.org/doc/draft-knodel-e2ee-definition/#:~:text=End%2Dto%2Dend%2Dencryption%20,integrity%20and%20authenticity%20for%20users.>

<sup>2</sup> For example, in *Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems*, the Center for Democracy & Technology defines end-to-end encryption as a service or app where the keys used to encrypt and decrypt data are known only to the senders and designated recipients of this data.



- **There are important choices to be made about what content policies apply in an end-to-end encrypted messaging environment.** Facebook’s Community Standards (which apply to Messenger) and Instagram’s Community Guidelines (which apply to Instagram DMs) play an important role in addressing potential adverse human rights impacts by setting the direction for what is and is not allowed on each platform. However, neither applies to WhatsApp, which has its own terms of service. There are two important questions to address: first, what content standards should apply to a private end-to-end encrypted messaging platform, and second, whether, in the context of cross-app communication, content standards should be consistent across the three messaging platforms.

On one hand, having the same set of standards would make content moderation decisions and user reporting easier, and would improve access to remediation. It would also avoid potential standards conflicts between users messaging across platforms.

On the other hand, creating one-size-fits-all standards might confuse users, who consent to a different kind of service when joining WhatsApp vs. Facebook vs. Instagram DMs and thus have different expectations. Having one set of standards could also be tricky because the different messaging platforms have different features that necessitate specific policies, such as peer-to-peer payments.

There is also the issue of enforcement to consider in developing policies: what kinds of policies would even be possible to effectively enforce in the context of end-to-end encrypted messaging, and how might enforcement differ across types of violative content, for example CSAM distribution vs. incitement to violence.

## 5.4 Product Factors

- **In an end-to-end encrypted environment, user reporting of problematic content and accounts is a critically important enforcement mechanism.** Unlike content shared to the public Facebook platform and Instagram, Meta will not be able to review messages for content that violates the Community Standards; this will mean that user reporting and tips from external sources (such as communications from law enforcement agencies, partners, and the media) will take on increased importance for identifying and addressing adverse human rights impacts.

*Meta will not be able to review messages for content that violates the Community Standards; this will mean that user reporting and tips from external sources (such as communications from law enforcement agencies, partners, and the media) will take on increased importance for identifying and addressing adverse human rights impacts.*

- **There are important human rights considerations when designing reporting channels and appeals mechanisms.** The ideal reporting channel would be designed to meet the needs of billions of rightsholders who could be anywhere in the world, who may speak any language, and who have a wide range of different digital capabilities. Additionally, given the likelihood that classifier-based approaches to identify platform abuse in end-to-end encrypted messaging will have error rates that result in users

being erroneously suspended, effective appeals mechanisms are also important. Given the challenges of scale, speed, and volume, it will be impossible for a “perfect” reporting and appeals channel to be created. However, the effectiveness criteria for nonjudicial grievance mechanisms contained in Principle 31 of the UNGPs (such as legitimacy, accessibility, predictability, equitability, and transparency) provides direction for a rights-based approach.

- **While user reporting is one way to enforce against problematic content and accounts, it does not prevent abuse from occurring.** In an end-to-end encrypted context, techniques such as identifying and utilizing behavioral signals to prevent harmful interactions, sending behavioral nudges, prompts, and warnings, and user education and guidance can all be used to prevent human rights harm by discouraging users from sharing problematic content or engaging in abusive behavior. These methods are still in their early stages, and work needs to be done to understand how best to implement them and how effective they are.
- **There is tension between the type of metadata collection and analysis required to mitigate many of the human rights risks of end-to-end encrypted messaging and the right to privacy.** Metadata collection and analysis of “behavioral signals” via classifiers will have increased importance for identifying misuse, high-risk behavior, and threat actors in an end-to-end encrypted environment. However, mass collection of metadata also presents privacy risks, which need to be carefully weighed and addressed—for example, by collecting and analyzing only the volume of metadata necessary for the task, and disclosing enough about these methods to allow for informed consent but not so much that bad actors are able to game the system. Some regulatory requirements, such as the EU e-Privacy Directive, may also limit or prohibit Meta’s ability to use metadata (and message content) to address human rights risks, illustrating the need to address this tension holistically.

*In an end-to-end encrypted context, techniques such as identifying and utilizing behavioral signals to prevent harmful interactions, sending behavioral nudges, prompts, and warnings, and user education and guidance can all be used to prevent human rights harm by discouraging users from sharing problematic content or engaging in abusive behavior.*

- **Using machine learning (ML) systems to detect and prevent problematic behavior and content is important for harm prevention and response at the scale of Meta, but on its own is not sufficient.** Although Meta continues to prioritize human review for many types of content and enforcement decisions, it is increasingly shifting toward automated removals as it improves its ML-based systems. ML can assist with risk and harm detection at scale. However, civil society organizations, researchers, and academics have shown that ML systems often struggle to account for context and nuance. Their outputs may be less accurate for vulnerable groups whose local languages and user behavior are less common, and are therefore not adequately reflected during the training and optimization of the system. This would result in new human rights concerns related to discrimination and equality. Although Meta should invest in improving the quality of its ML classifiers, adequate human review resources across geographic and linguistic contexts need to also be sufficiently allocated to enable nuanced analysis and mitigate the impacts of automated detection and enforcement errors.

- **Potential technical mitigations have been proposed for identifying and removing illegal content in an end-to-end encrypted messaging environment,<sup>3,4</sup> but the only approach proposed thus far that may not undermine cryptographic integrity is not technically feasible today.** Methods such as client-side scanning of a hash corpus, trained neural networks, and multiparty computation including partial or fully homomorphic encryption have all been suggested by some as solutions to enable messaging apps to identify, remove, and report content such as CSAM. They are often collectively referred to as “perceptual hashing” or “client-side scanning,” even though they can also be server-side.<sup>5</sup> Nearly all proposed client-side scanning approaches would undermine the cryptographic integrity of end-to-end encryption, which because it is so fundamental to privacy would constitute

*Potential technical mitigations have been proposed for identifying and removing illegal content in an end-to-end encrypted messaging environment, but the only approach proposed thus far that may not undermine cryptographic integrity is not technically feasible today.*

significant, disproportionate restrictions on a range of rights, and should therefore not be pursued (see Section 10).

Although homomorphic encryption fails to address the concerns of those who believe in a broader definition of end-to-end encryption (see above), it is the only approach proposed thus far that may not undermine cryptographic integrity, and advocates for homomorphic encryption argue it is the only approach that would not disproportionately infringe on the privacy and other rights of all users. However, homomorphic encryption is still nascent and theoretical, and is far too computationally intensive for even high-end mobile devices today. For example, Meta’s own research of a homomorphic encryption approach found that it would take around 20 million seconds to run (over seven months) on a single message. Further, any technical solution would need to work on low-end devices, which are used by a large number of Meta users, for it to be effective and respect the different circumstances of vulnerable groups. Nevertheless, research into these



- 3 See: Jonathan Mayer, Content Moderation for End-to-End Encrypted Messaging, Princeton University, October 6, 2019, [https://www.cs.princeton.edu/~jrmayer/papers/Content\\_Moderation\\_for\\_End-to-End\\_Encrypted\\_Messaging.pdf](https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf); Priyanka Singh and Hany Farid, Robust Homomorphic Image Hashing, Computer Vision Foundation Workshop, [http://openaccess.thecvf.com/content\\_CVPRW\\_2019/papers/Media%20Forensics/Singh\\_Robust\\_Homomorphic\\_Image\\_Hashing\\_CVPRW\\_2019\\_paper.pdf](http://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Singh_Robust_Homomorphic_Image_Hashing_CVPRW_2019_paper.pdf); Hany Farid, Opinion: Facebook’s Encryption Makes It Harder to Detect Child Abuse, Berkeley School of Information, October 25, 2019, <https://www.ischool.berkeley.edu/news/2019/opinion-facebooks-encryption-makes-it-harder-detect-child-abuse>.
- 4 The perspective of some experts proposing these approaches evolved during the course of this assessment. See: <https://www.washingtonpost.com/opinions/2021/08/19/apple-csam-abuse-encryption-security-privacy-dangerous/> and Identifying Harmful Media in End-to-End Encrypted Communication: Efficient Private Membership Computation.
- 5 Server-side means that the computation takes place on a web server, whereas client-side means the computation takes place on the user’s device.

methods is still in its early stages. Other novel approaches to client-side scanning that uphold cryptographic integrity may also be proposed, and computational power will likely eventually increase enough to enable such solutions.

*Even if homomorphic encryption and other proposed solutions were technically feasible and successfully maintained cryptographic integrity, they would still pose several other human rights risks*

- **Even if homomorphic encryption and other proposed solutions were technically feasible and successfully maintained cryptographic integrity, they would still pose several other human rights risks that would need to be addressed.** For example, if Meta starts detecting and reporting universally illegal content like CSAM, some governments are likely to exploit this capability by requiring Meta to block and report legitimate content they find objectionable, thereby infringing on the privacy and freedom of expression rights of users. It is noteworthy that even some prior proponents of homomorphic encryption have subsequently altered their perspective for this reason, concluding that their proposals would be too easily repurposed for surveillance and censorship.<sup>6</sup> In addition, these solutions are not foolproof; matching errors can occur, and bad actors may take advantage of the technical vulnerabilities of these solutions to circumvent or game the system. For these reasons, Meta and many other stakeholders argue that any form of content scanning should not be pursued for end-to-end encrypted messaging. It is also BSR's recommendation

that if the implementation of client-side scanning solely to detect CSAM—a legitimate aim—would likely result in a significant restriction of freedom of expression, privacy, and other rights, then client-side scanning should not be pursued. These issues are explored in greater detail in Sections 10 and 11.

- **Even with end-to-end encryption, the risks of malicious access to users' messages still exist.** The proliferation of corporate spyware has enabled governments around the world to gain remote access to target's smartphones and computers, allowing them to simply view end-to-end encrypted messages as if they were the user. For example, the NSO Group's Pegasus spyware has been discovered on the phones of journalists, activists, and political opponents around the world, from Mexico to Saudi Arabia.<sup>7</sup>
- **The human rights implications of cross-app communication are not fully known.** While this assessment touches on some elements of cross-app communication, such as the privacy implications of linked accounts and increased discoverability, an assessment to understand the full range of impacts has not been conducted. It will be important for the human rights impacts of cross-app communication to be further assessed, including their interaction with end-to-end encryption.

### 5.5 External Engagement Factors

- **Law enforcement concerns about not being able to access content shared on end-to-end encrypted messaging platforms should be considered in the broader context of a radically altered digital environment.** While a shift to end-to-end encryption may reduce law enforcement agency access to the content of some communications, it would be wrong to conclude that law enforcement agencies are faced with a net loss in capability overall. Trends such as the collection and analysis of significantly increased volumes of metadata, the value of behavioral

<sup>6</sup> See <https://www.washingtonpost.com/opinions/2021/08/19/apple-csam-abuse-encryption-security-privacy-dangerous/>.  
<sup>7</sup> <https://www.npr.org/2021/08/25/1027397544/nso-group-pegasus-spyware-mobile-israel> and <https://citizenlab.ca/tag/nso-group/>.

signals, and the increasing availability of artificial intelligence-based solutions run counter to the suggestion that law enforcement agencies will necessarily have less insight into the activities of bad actors than they did in the past. Innovative approaches can be deployed that may deliver similar or improved outcomes for law enforcement agencies, even in the context of end-to-end encryption. However, many law enforcement entities today lack the knowledge or the resources to take advantage of these approaches and are still relying on more traditional techniques.

- **Meta has a dilemma in deciding how to proactively collaborate with law enforcement agencies.** As described by the first pillar of the UNGPs, governments have a duty to protect human rights, and in an ideal world, governments would meet this duty in good faith. In these circumstances a case can be made that Meta proactively supporting law enforcement agencies' efforts to tackle legitimate crime in an end-to-end encrypted environment—for example, by helping them make better use of metadata analysis—can play an important role in Meta's responsibility to address human rights harm. However, in a growing number of cases, government intentions are not aligned with human rights or there is lack of rule of law,<sup>8</sup> making proactive collaboration with law enforcement agencies problematic in many contexts.
- **Meta will increasingly rely on user reporting, metadata, behavioral signals, and ML classifiers to address problematic content and interactions in end-to-end encrypted messaging.** However, engagement with law enforcement should consider that metadata analysis and behavioral signals cannot always provide the same level of certainty as access to actual message content may provide. When law enforcement agencies have access to message content via judicial order they may be able to discern definitively whether users are engaging in criminal activity. However, even the most accurate



machine learning systems that use metadata and behavioral patterns to identify criminal activity cannot be 100 percent certain. In the absence of user reporting, this information is likely still highly useful to law enforcement, but it is unclear to what extent it could be considered as evidence, or whether it should be provided to law enforcement at all. Such decisions could impact the right to a fair trial and freedom from arbitrary detention.

- **Proactive and productive public policy engagement on encryption is essential to address growing government attempts to ban or undermine end-to-end encryption.** The current binary “privacy-vs.-security” approach to advocacy that has dominated the encryption debate thus far has not proven effective, no matter how many cryptographers and security experts encryption defenders assemble in their ranks. In addition to proactive engagement with law enforcement, which should be done on a case-by-case basis in consideration of the human rights and rule of law context of the law enforcement entity, Meta will need to productively engage with other government officials to inform them of Meta's approach to assisting law enforcement, and all of the ways in which evidence and intelligence gathering can adapt to end-to-end encrypted contexts. Meta should also expand its outreach and engagement with civil society organizations and experts working in child protection to foster mutual understanding and advance solutions.

<sup>8</sup> Freedom House has documented the global trend of declining respect for democracy, human rights, and rule of law around the world: <https://freedomhouse.org/report/freedom-world>.

## On the relevance of Apple's announced rollout of new child safety features to this assessment.

In August 2021, Apple announced it would roll out several features in the US designed to protect children in messages, iCloud photos, and search. The announcement garnered both high-profile celebration and rebuke, and Apple ultimately slowed the rollout. Proponents applauded Apple for making strides in protecting children from sexual abuse and exploitation, while opponents raised concerns about the technical integrity of the features and argued that in today's volatile regulatory context, government entities are likely to abuse these capabilities beyond good faith actions to protect children by requiring companies to monitor, detect, remove, and report legitimate content. It is the latest case in the ongoing tension between protecting children from sexual abuse and exploitation online and protecting the privacy, security, and freedom of expression rights of all users.

Some of the initially announced features are relevant to this assessment; however, there are important differences as well.

At the time of writing, Apple's new child safety features are focused in three areas, and can be summed up as following:<sup>9</sup>

- **iMessage:** Apple will use on-device machine learning to identify sexually explicit images before they are viewed or sent by children who are enrolled in family accounts.
- **iCloud photos:** Before an image is stored in iCloud photos, an on-device matching process will be performed against a hash database of known child sexual abuse material. If 30 images are flagged as CSAM, Apple will conduct a human review to verify and make a report to the National Center for Missing and Exploited Children (NCMEC). This particular aspect of the Apple announcement has faced the most criticism from technologists and digital rights advocates.
- **Siri and Search:** Siri and search will provide additional resources to help children and parents stay safe online and will attempt to redirect users away from searches related to child sexual abuse.

Many of the tensions in the Apple case are shared by Meta in its planned expansion of end-to-end encrypted messaging. The use of encrypted messaging to facilitate the exploitation of children and share CSAM undetected is a key human rights risk identified in this assessment, and we explore the related tensions throughout. However, there are several important differences that make it challenging to simply copy-paste the debate about the Apple announcement to the challenges for Meta examined in this HRIA.

This assessment focuses specifically on end-to-end encrypted messaging, whereas Apple's product changes primarily involve cloud storage. There are some common challenges and risks associated with CSAM detection in both a cloud storage context and a messaging context. For example, both systems could be abused or "gamed" by bad actors to evade detection or trigger false matches, and both may open the door to governments requiring companies to monitor, detect, and report legitimate content, which could result in widespread surveillance and censorship.

However, Apple's proposal involves partially conducting the machine learning analysis of iMessages and the CSAM detection of iCloud photos on Apple devices at the operating system level. Apple devices are high-end, with higher levels of computational power. By contrast, Meta's messaging products are used on all manner of devices, both high- and low-end, and if Meta were to build similar features it could only be undertaken at the app level because Meta does not own or have access to the operating systems of devices.

In sum, the Apple announcement raised the profile of the debate about how to address child sexual abuse and exploitation issues in private and semi-private digital spaces while preserving broader privacy protections and other human rights of all users. This debate led to Apple ultimately pausing the rollout of these features. Important debates are taking place, and Sections 8.1, 10.2, and 11 of this assessment provide a deeper, nuanced, and rights-based analysis of these issues.

<sup>9</sup> [https://www.apple.com/child-safety/pdf/Security\\_Threat\\_Model\\_Review\\_of\\_Apple\\_Child\\_Safety\\_Features.pdf](https://www.apple.com/child-safety/pdf/Security_Threat_Model_Review_of_Apple_Child_Safety_Features.pdf).

# Potential Human Rights Impacts

This section assesses Meta’s expansion of end-to-end encryption against the human rights contained in the Universal Declaration of Human Rights and codified in international human rights instruments by evaluating each relevant human right for the risks and opportunities that could arise from end-to-end encrypted messaging.

As described in the methodology section, an HRIA typically considers how closely Meta would be connected to the potential human rights impacts using the UNGPs “cause/contribute/directly linked” framework. However, this framework is generally considered to only apply to adverse human rights impacts, rather than benefits. Because end-to-end encrypted messaging has so many human rights benefits, we have developed the following framework to illustrate how closely Meta would be connected to both the human rights harms and benefits. **This framework is not a part of the UNGPs**; however, this framework does inform our subsequent UNGPs-based analysis on “cause/contribute/directly linked” for potential adverse human rights impacts:

- **First-order impact:** Risks or opportunities that directly result from an action taken by Meta.
- **Second-order impact:** Risks or opportunities that directly result from actions taken by users or entities other than Meta.



- **Third-order impact:** Risks or opportunities that indirectly result from actions taken by users or entities other than Meta.

The intention of this framework is to illustrate **how Meta would be connected to both the human rights harms and benefits of end-to-end encrypted messaging**, which can help inform analysis of the leverage Meta has to minimize the potential harms and maximize the benefits.

This framework **does not imply that Meta has no responsibility to address second- or third-order human rights risks, or that it has no ability to do so**. The UNGPs are clear that companies must address all actual and potential human rights risks with which they are involved, including those they are “directly linked” to by their products and services.

This framework also **does not imply that Meta should prioritize addressing first-order risks**. The UNGPs are clear that companies must prioritize based on severity, and many of the most severe adverse human rights impacts associated with

end-to-end encrypted messaging are second- and third-order impacts.

The following table summarizes the potential human rights impacts of Meta’s shift to end-to-end encryption of all messaging platforms, as well as some of the related risks of the proposed cross-app communication of messaging platforms. It should be noted that Meta has already experienced many of these impacts on WhatsApp because it is already end-to-end encrypted.

Following the table we explore how the “cause/contribute/directly linked” framework can apply in the case of end-to-end encrypted messaging. We then discuss the most salient areas of human rights risks and opportunities in more detail.

Because this assessment reflects Meta’s expansion of end-to-end encryption across its messaging services rather than the deployment of end-to-end encryption for the first time, many of the adverse impacts detailed below already occur to some extent (i.e., they are actual impacts). The impact of expanding end-to-end encryption, therefore, will be

***Opportunities arising from deploying end-to-end encryption across all Meta’s messaging services are closely associated with the safety of the Meta platforms themselves when used as intended. By contrast, the risks tend to be associated with the actions of bad actors disregarding terms of service, violating the law, and adversely impacting the rights of others.***

the potential expansion of these adverse impacts across all of Meta’s messaging platforms.

It is noteworthy that opportunities arising from deploying end-to-end encryption across all Meta’s messaging services are closely associated with the safety of the Meta platforms themselves when used as intended. By contrast, the risks tend to be associated with the actions of bad actors disregarding terms of service, violating the law, and adversely impacting the rights of others. In other words, the human rights opportunities of end-to-end encrypted messaging are often first-order effects, whereas the potential human rights harms are often second- or third-order effects. This has significant implications for how Meta can address potentially adverse human rights impacts, and which other actors, such as individuals and governments, also have a role to play.

It is also important to note that both users and nonusers benefit from the human rights opportunities of end-to-end encrypted messaging and suffer from the human rights harms—meaning users aren’t the only rightsholders involved. For example, the information and activities of nonusers may be protected by end-to-end encryption (such as participants in a peaceful protest organized via WhatsApp), and because users can use end-to-end encrypted messaging in ways that harm nonusers (such as via the fomenting of violence against members of a certain ethnic group).

Further, although there are significant child rights risks arising from the expansion of end-to-end encryption, which we detail below, it is important to note that children will also benefit from the human rights opportunities listed here, such as increased privacy, greater opportunities for freedom of opinion and expression, and physical safety.

Later in this report we discuss how some of the potential mitigations for the human rights risks can themselves have negative human rights implications; this is discussed in further detail in the recommendations section.



## Actual and Potential Human Rights Impacts of Meta's Shift to End-to-End Encryption of All Messaging Platforms.

Right	Relevant Articles	Risks and Opportunities
<b>Rights to Equality and Nondiscrimination</b>	UDHR Article 2, 23 ICCPR Articles 2, 6 CRC Article 2 ICERD Article 5 UNDRIP Articles 13, 15, 16, 17 CRPD Article 27 CEDAW Article 11 ICESCR Article 6 ILO C100, C111	<p><b>Opportunities: (1st-order impacts)</b></p> <ul style="list-style-type: none"> <li>• End-to-end encrypted communications reach all users of Meta messaging platforms and not just those who know or care about encryption enough to opt in to end-to-end encrypted platforms.</li> <li>• End-to-end encrypted messaging could protect communications of partners in places where certain groups are legally or socially restricted from marriage, such as LGBTQIA+ persons and ethnic groups.</li> </ul> <p><b>Risks (2nd-and 3rd-order impacts)</b></p> <ul style="list-style-type: none"> <li>• Content that intends to harass users, including children, based on characteristics such as gender, religion, ethnicity, or political views may be shared on messaging platforms, but not reported and / or removed (2nd-order impact).</li> <li>• Misinformation and disinformation that is intended to promote discrimination may be posted on messaging platforms, but not reported and / or removed (2nd-order impact).</li> <li>• Use of behavioral signals and metadata analysis may result in law enforcement actions that are discriminatory in nature (3rd-order impact).</li> </ul>

Right	Relevant Articles	Risks and Opportunities
<b>Bodily Security Rights</b>	UDHR Article 3 CRPD Articles 10, 14 CRPD Article 10 CRC Article 6 ICCPR Articles 6, 9, 20 UNDRIP Article 7	<p><b>Opportunities: (1st-order impacts)</b></p> <ul style="list-style-type: none"> <li>The privacy protections of Meta’s end-to-end encrypted messaging platforms keep people safe from bad actors who would use their message content to cause them bodily harm or detain them arbitrarily. This is particularly true for vulnerable groups. Examples include keeping human rights defenders, journalists, and political dissidents safe from authoritarian governments, keeping women safe from spying partners or family, and keeping members of the LGBTQIA+ community safe from unfriendly governments or citizens.</li> </ul> <p><b>Risks: (2nd- and 3rd-order impacts)</b></p> <ul style="list-style-type: none"> <li>Bad actors may use Meta’s end-to-end encrypted messaging platforms undetected to traffic adults and children, through the coordination of trafficking and forced prostitution. Bad actors may also use messaging for predatory grooming that can lead to child sexual exploitation. These may result in bodily harm to both adults and children who are trafficked and / or abused (3rd-order impact).</li> <li>Cross-app communication of messaging platforms may make it easier for users to find people on other platforms. This increased “discoverability” could make it easier for bad actors to identify and contact victims (2nd-order impact).</li> <li>Bad actors may use Meta’s end-to-end encrypted messaging platforms undetected to share live video of child abuse, including sexual abuse and exploitation, via end-to-end encrypted video calls (2nd-order impact).</li> <li>Users may use end-to-end encrypted messaging platforms in a way that incites or encourages children to commit suicide (2nd-order impacts).</li> <li>Terrorists may use Meta’s end-to-end encrypted messaging platforms undetected to plan a terrorist attack that results in injuries and / or deaths (3rd-order impact).</li> <li>Criminals may use Meta’s end-to-end encrypted messaging platforms undetected to sell illicit goods and/or carry out crimes that result in injuries and/or deaths (3rd order impact).</li> <li>Users may share hate speech, misinformation, and disinformation to large groups on Meta’s end-to-end encrypted messaging platforms that achieve viral reach. This content may exacerbate existing social tensions and spark real-world violence that results in injuries and/or deaths (3rd-order impact).</li> <li>Coordinated behavior may incite violence or hostility against certain groups, resulting in harm to bodily security (2nd- and 3rd-order impact).</li> </ul>

Right	Relevant Articles	Risks and Opportunities
<p><b>Freedom From Slavery</b></p>	<p>UDHR Article 4 ICCPR Article 8 CRC Article 35 CEDAW Article 6 ILO C29, C105 ILO C138, C182</p>	<p><b>Risks: (3rd-order impacts)</b></p> <ul style="list-style-type: none"> <li>• Meta’s end-to-end encrypted messaging platforms may be used by human traffickers to facilitate trafficking that results in slavery.</li> </ul>
<p><b>Freedom From Torture, Degrading Treatment, or Punishment</b></p> <p>Including right to freedom from exploitation, violence, and abuse</p> <p>Including protection of children from physical or mental violence, injury or abuse, neglect or negligent treatment, maltreatment or exploitation, including sexual exploitation and abuse</p>	<p>UDHR Article 5 ICCPR Article 7 CRPD Articles 15, 16, 37 CATCIDTP Articles 13, 14 CRC Article 19, 34</p>	<p><b>Opportunities (1st-order impacts)</b></p> <ul style="list-style-type: none"> <li>• The privacy protections of Meta’s end-to-end encrypted messaging platforms keep people safe from violence inflicted by bad actors who would use their message content to cause them harm. This is particularly true for vulnerable groups. Examples include keeping human rights defenders, journalists, and political dissidents safe from authoritarian governments, keeping women safe from spying husbands or family, and keeping members of the LGBTQIA+ community safe from unfriendly governments or citizens.</li> </ul> <p><b>Risks: (2nd- and 3rd-order impacts)</b></p> <ul style="list-style-type: none"> <li>• Meta’s end-to-end encrypted messaging platforms may be used to facilitate the degrading treatment of trafficked/enslaved people (3rd-order impact).</li> <li>• Bad actors may use Meta’s end-to-end encrypted messaging platforms to facilitate activities that involve physical and mental violence against, and exploitation of, children (2nd- and 3rd-order impacts).</li> <li>• Bad actors may use Meta’s end-to-end encrypted messaging platforms undetected to share live video of children being sexually abused and exploited (2nd-order impact).</li> <li>• Users may share child sexual abuse material (CSAM) undetected on Meta’s end-to-end encrypted messaging platforms, resulting in “mental violence”/ psychological harm to the children involved who are aware this material is being shared (3rd-order impact).</li> <li>• Cross-app communication of messaging platforms may make it easier for users to find people on other platforms. This increased “discoverability” could make it easier for bad actors to identify and contact victims (2nd-order impact).</li> </ul>

Right	Relevant Articles	Risks and Opportunities
<b>Right to Remedy</b>	UDHR Article 8 ICCPR, Article 2 UNDRIP Article 40	<p><b>Opportunities (2nd-order impacts):</b></p> <ul style="list-style-type: none"> <li>The privacy protections of end-to-end encrypted messaging may increase the likelihood and security of whistleblowing, reporting, and exposing human rights violations, which thereby increases the likelihood of remediation.</li> </ul> <p><b>Risks (1st-order impacts):</b></p> <ul style="list-style-type: none"> <li>The way in which Meta implements user reporting of problematic content and/or abusive accounts in an end-to-end encrypted environment may impact the right to remedy if Meta is unable to effectively respond to reports and take appropriate action against users who violate Community Standards.</li> </ul>
<b>Freedom From Arbitrary Arrest and Exile</b>	UDHR Article 9 ICCPR Article 9	<p><b>Opportunities: (1st-order impacts)</b></p> <ul style="list-style-type: none"> <li>The privacy protections of Meta’s end-to-end encrypted messaging platforms keep people safe from arbitrary arrest based on the content of their messages. This is particularly true for certain vulnerable groups in countries without adequate rule of law, including human rights defenders, journalists, political dissidents, and members of the LGBTQIA+ community.</li> </ul>
<b>Right to Privacy</b>	UDHR Article 12 ICCPR Article 17 CRPD Article 22 CRC Article 16 UNDRIP Articles 12, 31	<p><b>Opportunities: (1st-order impacts)</b></p> <ul style="list-style-type: none"> <li>By ensuring that only the parties in a given conversation can see the content of messages, end-to-end encryption protects the privacy rights of users. By rolling out end-to-end encryption across all of its messaging platforms, Meta will be extending these privacy protections to all users.</li> <li>Privacy is an “enabling right” – the privacy protections of end-to-end encrypted messaging directly enable users to freely exercise many other human rights.</li> </ul> <p><b>Risks (1st-order impacts):</b></p> <ul style="list-style-type: none"> <li>End-to-end encrypted messaging may be used to share content that violates people’s (including children’s) privacy each time it is shared, such as nonconsensual intimate images and CSAM.</li> <li>Making all messaging platforms capable of cross-communication may enable people to find users on different platforms, increasing “discoverability.” This places at risk the privacy of users who do not have accounts on all platforms and/or do not wish to be discoverable across platforms. This could cause particular harm to users who maintain anonymous accounts and do not wish for their identities to be known. This includes human rights defenders and journalists who maintain anonymous accounts in order to share information without fear of retribution. Although account linking will be optional, the way in which the user decision is constructed will impact to what extent this risk is mitigated. Existing user controls on who can contact them will also be key.</li> <li>Making all messaging platforms capable of cross-communication increases the amount of user data Meta has access to and results in increased privacy risks through linking of accounts and data.</li> </ul>

Right	Relevant Articles	Risks and Opportunities
<b>Freedom of Movement</b>	UDHR Article 13 CRPD Article 18 ICCPR Article 12	<p><b>Opportunities: (1st-order impacts)</b></p> <ul style="list-style-type: none"> <li>The privacy protections of end-to-end encrypted messaging enables freedom of movement of those for whom access to the content of their communications could be used to restrict their movement, for example protest and political organizing. This is especially true for vulnerable groups who are more likely to have their rights restricted.</li> </ul>
<b>Freedom of Thought, Conscience, and Religion</b>	UDHR Article 18 ICCPR Article 18 UNDRIP Article 12 CRC Article 14	<p><b>Opportunities: (1st-order impacts)</b></p> <ul style="list-style-type: none"> <li>The privacy protections of end-to-end encrypted messaging enable people to freely and safely share beliefs and practice their religion in places where that right is restricted.</li> </ul> <p><b>Risks (3rd-order impacts):</b></p> <ul style="list-style-type: none"> <li>Users may share hate speech targeting members of a certain religious group on Meta's end-to-end encrypted messaging platforms. This speech may encourage users to harass and harm members of that religious group in such a way that they are unable to gather and practice their religion openly without fear of violence.</li> </ul>
<b>Freedom of Opinion, Expression, and Information</b>	UDHR Article 19 ICCPR, Article 19 CRPD Articles 7, 21 CRC Article 12, 13 UNDRIP Articles 13, 16	<p><b>Opportunities: (1st-order impacts)</b></p> <ul style="list-style-type: none"> <li>The privacy protections of end-to-end encrypted messaging enable people to freely express themselves and share, receive, and access information without fear of retribution in places and contexts where freedom of expression and opinion and access to information are restricted.</li> </ul> <p><b>Risks (3rd-order impacts):</b></p> <ul style="list-style-type: none"> <li>Because end-to-end encryption makes it challenging to detect the spread of hate speech and misinformation in messaging, users in group messages may increasingly self-censor for fear of being targeted.</li> </ul>
<b>Right of Children to be Protected From Harmful Information</b>	CRC Article 17	<p><b>Risks (2nd-order impacts)</b></p> <ul style="list-style-type: none"> <li>Meta's end-to-end encrypted messaging platforms may be used to propagate material that is harmful to children undetected. This includes material shared in messages among children, and in messages between children and adults.</li> </ul>
<b>Freedom of Assembly and Association</b>	UDHR Article 20 ICCPR Article 21 ICESCR Article 8 CRC Article 15 ILO C87, C98	<p><b>Opportunities: (1st-order impacts)</b></p> <ul style="list-style-type: none"> <li>The privacy protections of end-to-end encrypted messaging enable people to organize both physical and virtual gatherings in places and contexts where freedom of assembly and association is restricted.</li> </ul> <p><b>Risks (3rd-order impacts):</b></p> <ul style="list-style-type: none"> <li>Because end-to-end encryption makes it challenging to detect hate speech and misinformation in messaging, people targeted by hate speech may not feel they can safely gather with others or otherwise participate freely in public life.</li> </ul>

Right	Relevant Articles	Risks and Opportunities
<b>Right to Participate in Government</b>	UDHR Article 21 ICCPR Article 25 CEDAW Article 7 CRPD Article 29 UNDRIP Article 18	<p><b>Opportunities: (1st-order impacts)</b></p> <ul style="list-style-type: none"> <li>The privacy protections of end-to-end encrypted messaging enable people to freely and safely discuss and facilitate participation in government in situations where there are attempts to interfere with free and fair elections.</li> </ul> <p><b>Risks (3rd-order impacts):</b></p> <ul style="list-style-type: none"> <li>Because end-to-end encryption makes it challenging to detect hate speech and misinformation in messaging, people targeted by hate speech may not feel they can safely participate in government, including voting and attending political events.</li> </ul>
<b>Right to Work, Equal Pay, and Fair Wages</b>	UDHR Article 23 ICESCR Articles 6, 7 CRPD Article 27 CEDAW Article 11 CRPD Article 27 ILO C100, C190 UNDRIP Article 17	<p><b>Opportunities: (2nd-order impacts)</b></p> <ul style="list-style-type: none"> <li>By helping enable freedom of association, privacy protections of end-to-end encrypted messaging can enable and protect labor union communication, recruitment, and activity in places and contexts where labor rights are restricted.</li> </ul>
<b>Right to an Adequate Standard of Living</b>	UDHR Article 26 ICESCR Article 11 CRC 27	<p><b>Opportunities (2nd-order impacts):</b></p> <ul style="list-style-type: none"> <li>End-to-end encrypted messaging could facilitate more seamless, secure forms of e-commerce and digital payment / mobile money. If adequately geared toward addressing the commerce and banking needs of low-income groups, this could increase standard of living.</li> </ul>
<b>Right to Participate in Cultural Life</b>	UDHR Article 27 ICCPR Article 27 CRPD Article 30 ICESCR Article 15 UNDRIP Articles 11, 13, 31	<p><b>Opportunities (1st- and 2nd-order impacts):</b></p> <ul style="list-style-type: none"> <li>The privacy protections of end-to-end encrypted messaging enable community members to maintain cultural ties in contexts where their culture is socially or legally repressed (1st-order impact).</li> <li>By being present in widely used messaging platforms, end-to-end encryption would enable more people to enjoy the benefits of scientific progress (2nd-order impact).</li> </ul> <p><b>Risks (3rd-order impacts):</b></p> <ul style="list-style-type: none"> <li>Because end-to-end encryption makes it challenging to detect hate speech and misinformation in messaging, people targeted by hate speech may not feel they can safely participate in cultural life.</li> </ul>
<b>Right to Benefit from Scientific Advancement</b>	UDHR Article 27 ICESCR Article 15	<p><b>Opportunities (1st-order impacts):</b></p> <ul style="list-style-type: none"> <li>Because Meta's family of apps are used by so many people around the world (over 2.8 billion), by extending end-to-end encryption across all its messaging services Meta is providing access to a key technology for protecting privacy and security. This ensures end-to-end encrypted messaging services are broadly accessible around the world, and not just to those with the knowledge or wherewithal to seek out specific tools.</li> </ul>

## 6.1 Attribution and End-to-End Encryption

When conducting HRIAs, BSR considers how closely the company would be connected to the human rights impact and the appropriate action that results using the following definitions outlined in the UNGPs:

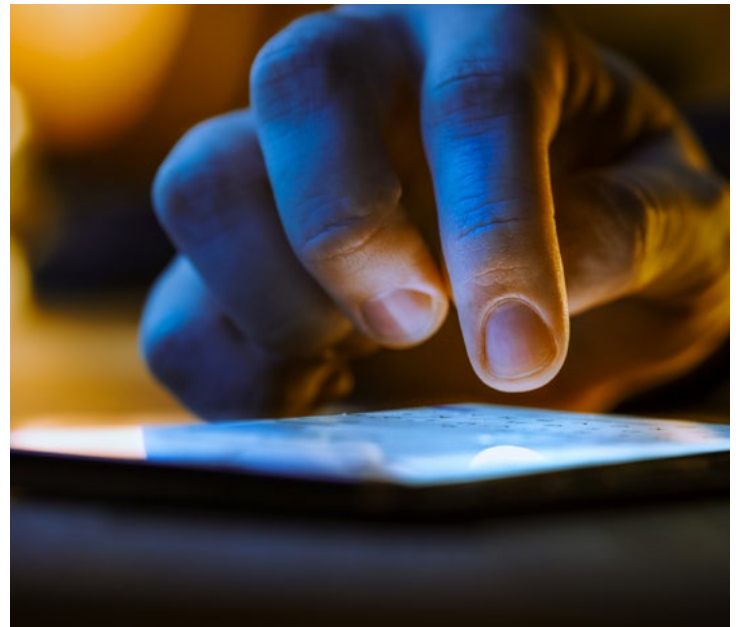
- **“Caused” the impact**—The company should take the necessary steps to cease or prevent the impact.
- **“Contributed” to the impact**—The company should take the necessary steps to cease or prevent its contribution and use its leverage to mitigate any remaining impact to the greatest extent possible.
- **“Directly linked”** to the impact through its products, services, or operations arising from its business relationships, including users<sup>1</sup> —The company should determine action based on factors such as the extent of leverage over the entity concerned and the severity of the harm.

By illustrating how a company is connected to human rights harm, the “cause, contribute, directly linked” framework suggests the appropriate action the company should take.

The UNGPs also state that when companies are found to have “caused” or “contributed to” adverse impacts, they should provide for or cooperate in remediation processes. A company that is “directly linked” to harm by their business relationships is not required to provide or cooperate in remediation, though it may take a role in doing so.

Applying the “cause, contribute, directly linked” framework to social media platforms and technology products is challenging due to the complex ways in which platforms interact with, enable, and amplify human behavior.

In the context of this HRIA, it is clear that **end-to-end encryption in and of itself does not “cause”**



**human rights harm.** Rather, most of the adverse human rights impacts that could result from the shift to end-to-encryption would occur due to other individuals or entities using end-to-end encrypted messaging to harm the human rights of others.

On its face, this would lead to an attribution of being “directly linked” to an adverse human rights impact. However, there are a few situations in which companies may be considered to be “contributing to” adverse human rights impacts via the actions of third parties.

According to the most recent literature, a technology company is more likely to be considered “contributing to” adverse human rights impacts if it takes actions (or fails to take actions) that:

- **Facilitate or enable** another entity to “cause” an adverse impact, where a company’s actions add to the conditions that **make it possible** for use of a product by a third party to “cause” harm.
- **Incentivize or motivate** another entity to “cause” an adverse impact, where a company’s actions **make it more likely** that a product or service will be used in ways that “cause” harm.<sup>2</sup>

<sup>1</sup> The relationship between technology companies and end users is generally considered a “business relationship” when interpreting the UNGPs for technology companies—and for this reason, technology companies can be “directly linked” to human rights harms carried out by end users, whether they be individuals or entities. For example, see this paper from the UN B-Tech Project, which has been charged with providing authoritative guidance on implementing the UNGPs in technology sector: <https://www.ohchr.org/Documents/Issues/Business/B-Tech/taking-action-address-human-rights-risks.pdf>.

<sup>2</sup> <https://www.ohchr.org/Documents/Issues/Business/B-Tech/taking-action-address-human-rights-risks.pdf>.



*Assuming Meta does adopt appropriate mitigation measures—such as the recommendations contained in this assessment—then BSR considers Meta to be directly linked to (rather than contributing to) the potential adverse human rights impacts associated with the expansion of end-to-end encryption.*

A key question arising in this HRIA is therefore **whether the deployment of end-to-end encryption across Meta’s messaging platforms would facilitate, enable, incentivize, or motivate others to “cause” harm.**

In and of itself, end-to-end encryption does not “contribute to” (i.e., enable, facilitate, incentivize, or motivate) harm because nearly all the adverse human rights impacts that could be attributed to end-to-end encryption already occur in non-end-

to-end encrypted messaging. Rather, the impact of end-to-end encryption is to potentially make this harm more difficult to detect.

However, it is reasonably foreseeable that making some harms more difficult to detect would increase the volume of adverse human rights impacts in end-to-end encrypted messaging. If this were to happen in reality, then **Meta would be “contributing to” those harm if reasonable mitigation measures are not put in place.**

For example, the distribution of CSAM is a violation of the privacy rights of the children involved and revictimizes them every time it is shared. CSAM is more difficult to detect in an end-to-end encrypted environment, and it is reasonably foreseeable that it may become more common if other measures to prevent, detect, and address CSAM are not put in place. For this reason, Meta may be considered as “contributing to” the harm, but only if it implements end-to-end encryption without adopting reasonable mitigation measures.

However, **assuming Meta does adopt reasonable mitigation measures—such as the recommendations contained in this assessment—then BSR considers Meta to be “directly linked” to (rather than “contributing to”) the potential adverse human rights impacts**



**associated with the expansion of end-to-end encryption.** Reasonable mitigation measures are described in more detail in the recommendations section, but include:

- Continuing to detect CSAM in unencrypted content such as profile photos and group photos.
- Using behavioral signals, public platform information, user reports, and metadata to identify and interrupt CSAM distribution groups.
- Working with law enforcement agencies in collaborative, rights-respecting efforts to identify users creating, distributing, and viewing CSAM.
- If and when technically feasible, detecting CSAM in end-to-end encrypted messaging using a method (such as homomorphic encryption) that does not undermine cryptographic integrity, but only after a review of the potential for adverse impacts on privacy, freedom of expression, and other rights, and a conclusion that those adverse impacts can be adequately addressed (see Section 11 for a discussion of the human rights risks associated with content detection in messaging).

Absent reasonable mitigation measures, Meta would be responsible for providing or cooperating in remediation for harms, such as by funding rehabilitation groups who work with CSAM victims

or through financial compensation for victims.

BSR notes that WhatsApp is actively pursuing mitigation measures to address potential harm to children in the context of end-to-end encryption. While WhatsApp does not and cannot review encrypted content, it currently relies on unencrypted information (e.g., group profile information and photos, and user reports) to ban hundreds of thousands of accounts per month suspected of sharing CSAM.

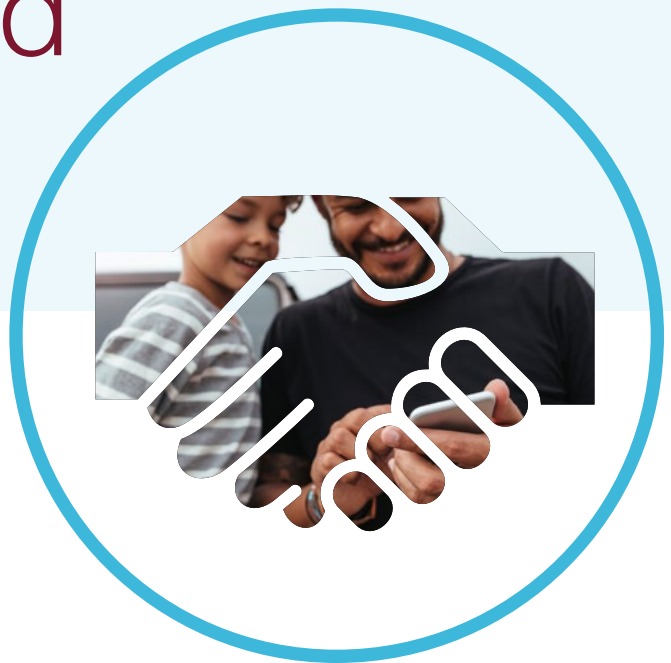
However, it is also important to note that **a decision not to implement end-to-end encryption would also more closely connect Meta to human rights harm. If Meta were to choose not to implement end-to-end encryption across its messaging platforms in the emerging era of increased surveillance, hacking, and cyberattacks, then it could be considered to be “contributing to” many adverse human rights impacts due to a failure to protect the privacy of user communications.**

In short, **the most effective way for Meta to ensure that it is not considered “contributing to” harms when deploying end-to-end encryption is by putting in place effective measures to address the potential adverse human rights impacts identified in this HRIA.** This can be achieved, for example, by implementing BSR’s recommendations in Section 12.

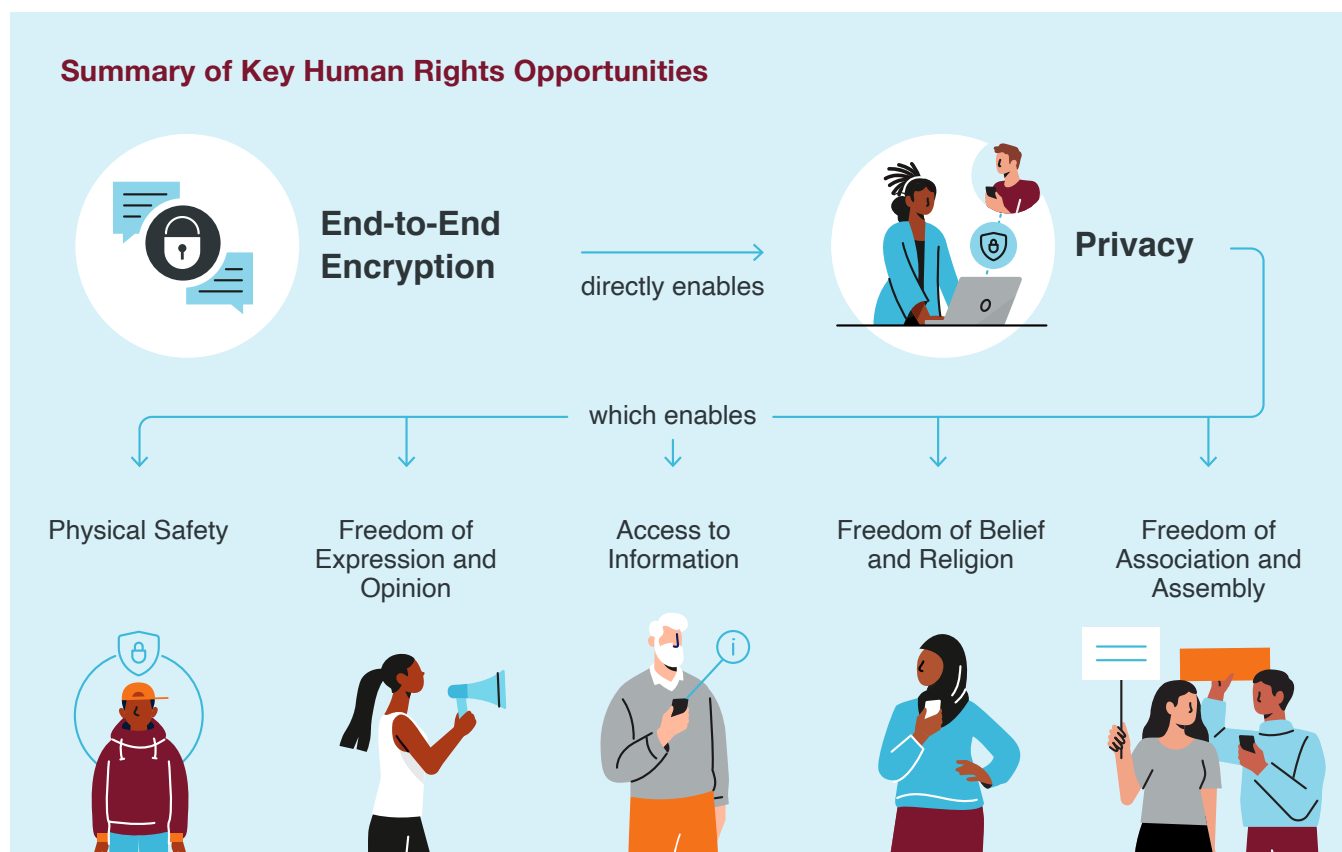
# Exploring the Key Rights Opportunities of Meta's Expansion of End-to-End Encryption

Although HRIAs tend to focus on human rights risks rather than opportunities, end-to-end encrypted messaging has numerous human rights benefits that are important to discuss. In this case, the expansion of end-to end encryption will directly result in the increased realization of a range of human rights. Conversely, the absence of ubiquitous end-to-end encryption of messaging platforms is likely to result in human rights risks.

When examining the human rights benefits it is important to note that end-to-end encrypted messaging directly enables the exercise of numerous human rights. In particular, because end-to-end encryption protects the right to privacy, it also enables freedom of opinion,



*Because end-to-end encryption protects the right to privacy, it also enables freedom of opinion, expression, movement, association, and assembly, as well as protects the physical safety of vulnerable users and other rightsholders*



expression, movement, association, and assembly (among many other rights—see Section 6), as well as protects the physical safety of vulnerable users and other rightsholders. These positive human rights impacts are expansive, benefiting all of the over 2.8 billion Meta users. However, the benefits are particularly important for members of vulnerable groups and users who live in contexts with a higher risk of human rights harm.

Additionally, because the key opportunities of end-to-end encrypted messaging are the direct enabling and protection of human rights, they are mutually reinforcing. Human rights are interrelated and interdependent, and that is reflected in the ways in which users and other rightsholders experience the human rights protections provided by end-to-end encrypted messaging. The main human rights opportunities are:

- **Opportunity 1: Enabling Privacy and Its Knock-On Benefits**

Privacy both enables and reinforces other human rights. When the right to privacy is respected, people can more freely exercise other rights that depend on privacy. This includes free expression, opinion, association, movement, religion, and belief, among many others.

- **Opportunity 2: Enabling Physical Safety**

For many vulnerable communities, end-to-end encrypted messaging does not just protect their privacy and enable free expression and association, it is also vital to their physical safety.

- **Opportunity 3: Enabling Free Expression and Opinion, Belief and Religion, Association and Assembly, and Access to Information**

By ensuring the privacy of communications, end-to-end encrypted messaging enables people to freely express themselves, access information, and assemble without fear of retribution.

## 7.1 Enabling Privacy

By ensuring that conversations remain only between intended participants, and allowing participants to verify this, the very nature of end-to-end encrypted messaging protects privacy. This has been underscored by David Kaye, former UN Special Rapporteur on Freedom of Expression, who stated, “Encryption and anonymity provide individuals and groups with a zone of privacy online to hold opinions and exercise freedom of expression without arbitrary and unlawful interference or attacks.”<sup>1</sup>

The privacy enabled by end-to-end encrypted messaging not only protects users and other rightsholders from the threat of bad state actors and other external groups, but it also protects against the risk of abuse within Meta. Currently, numerous Meta employees have access to unencrypted messages shared across the platforms as part of the business of running the platform, and some employees may be tempted to abuse this access to spy on individuals of interest, whether for personal reasons<sup>2</sup> or because they have been recruited by a state actor to use their access for espionage. The recent exposure of Saudi government espionage at Twitter via Saudi employees revealed the likelihood of insider threats inside major technology companies.<sup>3</sup> Deploying end-to-end encryption would help ensure that Meta employees cannot violate the communications privacy of users.

Because Messenger only offers opt-in end-to-end encryption and Instagram DMs is currently only in the testing phase, the private conversations of billions of users are currently unprotected. By deploying end-to-end encryption across all messaging platforms, Meta will be extending the privacy protections of end-to-end encryption—currently provided by WhatsApp—to all Facebook Messenger and Instagram DMs users around the world.

### IMPACT FACTORS

**While all users can benefit from privacy and its follow-on effects, the privacy benefits of end-to-end encrypted messaging are particularly important for members of vulnerable groups. For example, protecting communications between members of the LGBTQIA+ community from those who would do them harm, or allowing victims of trafficking to securely contact support services.**

**This vulnerability is both dependent upon and amplified by geographic context. For example, in countries where surveillance is widespread and dissent is not tolerated and dissenters are jailed or worse, end-to-end encrypted messaging is particularly important for human rights activists and journalists to be able to conduct their work.**

## 7.2 Enabling Physical Safety

For many vulnerable communities, end-to-end encrypted messaging is vital to their physical safety. The consequences of malicious actors intercepting the communications of a human rights activist or a journalist investigating corruption could be arbitrary detention, bodily harm, torture or other cruel or inhumane treatment, or even death.

In this way, end-to-end encrypted messaging can enable the protection of some of the most fundamental human rights that exist, and that cannot be restricted or derogated by governments under any circumstances. Extending end-to-end encryptions across messaging platforms will provide vital safety protections for vulnerable users and other rightsholders around the world. However, it is also important to note that end-to-end encrypted messaging comes with physical safety risks too, such as when users use it to carry out terrorist attacks or facilitate the sexual exploitation and abuse of children.

1 [https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session29/Documents/A.HRC.29.32\\_AEV.doc](https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session29/Documents/A.HRC.29.32_AEV.doc).

2 [https://www.vice.com/en\\_us/article/bjp9zv/facebook-employees-look-at-user-data](https://www.vice.com/en_us/article/bjp9zv/facebook-employees-look-at-user-data).

3 <https://www.nytimes.com/2019/11/06/technology/twitter-saudi-arabia-spies.html>.

## IMPACT FACTORS

**Risks to physical safety are most prominent for members of vulnerable groups, particularly those who frequently face physical violence, such as victims of trafficking and domestic violence, women activists, children, and members of the LGBTQIA+ community. These risks are exacerbated by geographic context, particularly in countries that do not fully respect human rights or the rule of law and where human rights abuses by security forces is common, as well as in countries with a history of sectarian violence.**

### 7.3 Enabling Free Opinion and Expression, Belief and Religion, Association and Assembly, and Access to Information

Privacy is a fundamental human right not just on its own—it is a foundation for the exercise of many other human rights—and both enables and reinforces other human rights.<sup>4</sup> By ensuring the

privacy of communications, end-to-end encrypted messaging enables people to freely express themselves, access information, and assemble without fear of retribution. It also enables people to freely practice their religion in community with others in places where that right is restricted. David Kaye has specifically written about how end-to-end encryption both protects and enables free expression and opinion, particularly for people in places where those rights are restricted, saying, “In environments of prevalent censorship, individuals may be forced to rely on end-to-end encryption and anonymity in order to circumvent restrictions and exercise the right to seek, receive and impart information.”<sup>5</sup>

While most users in democratic countries where free expression is protected experience fewer limitations on their ability to express themselves in unencrypted environments, vulnerable users and users in high human rights risk countries can face significant limitations. In such contexts, users may choose to self-censor to avoid the risks that come from speaking freely.



4 See General Assembly resolution 68/167 (<http://www.ohchr.org/en/issues/digitalage/pages/digitalageindex.aspx>), A/HRC/13/37 (<http://www2.ohchr.org/english/bodies/hrcouncil/docs/13session/a-hrc-13-37.pdf>) and Human Rights Council resolution 20/8 (<https://documents-dds-ny.un.org/doc/RESOLUTION/GEN/G12/153/25/PDF/G1215325.pdf?OpenElement>)

5 [https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session29/Documents/A.HRC.29.32\\_AEV.doc](https://www.ohchr.org/EN/HRBodies/HRC/RegularSessions/Session29/Documents/A.HRC.29.32_AEV.doc)

The right to free association and assembly provides for the right to gather both publicly and privately to pursue common interests, including everything from participating in a protest march to seeing a theater show or participating in a group discussion. End-to-end encrypted messaging enables people to exercise their rights to free association and assembly both virtually and in-person. Groups organize through messaging applications, whether for a casual meetup or a march for women's rights. Sometimes this organizing results in in-person gatherings, but the mere act of exchanging end-to-end encrypted messages also enables people to associate and assemble virtually.

WhatsApp is widely used as a space for people to gather, in both small and large groups. Although group messages make up only 10 percent of WhatsApp conversations today, and the majority consist of groups smaller than 10 people, group messages have provided users with vital channels to connect with each other and pursue common interests, whether that interest be simply sharing information and connecting with people or toward a more defined group objective.

Deploying end-to-end encryption and cross-app communication across all Meta messaging platforms enables the expansion of this right beyond just WhatsApp. For most users in democratic countries, these benefits are likely to be felt in the form of greater ease of connecting with others. For vulnerable users and rightsholders in at-risk geographic contexts where they may be restricted from gathering physically, the freedom of association benefits of end-to-end encrypted messaging are more pronounced.

## IMPACT FACTORS

**The opportunity for protecting and enabling free expression, opinion, belief, religion, assembly, and association are particularly important for members of vulnerable groups who are often prevented from expressing themselves freely or are attacked when they do. Typically vulnerable groups include, but are not limited to, human rights defenders, journalists, political dissidents, environmental and community activists, women, children, members of ethnic and religious minorities, indigenous groups, the elderly, members of the LGBTQIA+ community, and those who are illiterate or digitally illiterate. However, vulnerability depends on context, and someone who may be powerful in one context may be vulnerable in another. These benefits of end-to-end encrypted messaging are particularly pronounced in countries where such rights are restricted across the population, where every citizen can face retribution or censorship and thus widely benefit from the protections of end-to-end encrypted messaging.**

# Exploring the Key Human Rights Risks of Meta's Expansion of End-to-End Encryption

Expanding end-to-end encryption will address many human rights risks associated with the absence of ubiquitous end-to-end encryption today, but it will also increase the salience of other human rights risks.

In contrast to the opportunities—where end-to-end encrypted messaging directly enables numerous human rights—the human rights risks of Meta's expansion of end-to-end encryption are largely associated with the actions of bad actors using an end-to-end encrypted environment to disregard terms of service, violate the law, and adversely impact the rights of others in ways that are challenging to detect. This means that end-to-end encryption itself does not directly “cause” adverse human rights impacts, though it may “contribute to” or be “directly linked” to them.

It is important to note that these risks exist across all messaging platforms; end-to-end encryption simply makes them more difficult to detect. However, some risks may become more



*Expanding end-to-end encryption will address many human rights risks associated with the absence of ubiquitous end-to-end encryption today, but it will also increase the salience of other human rights risks.*

prevalent as bad actors seek to exploit the privacy protections of end-to-end encrypted messaging. It is also important to note that compared to the opportunities, which extend to all users of Meta's messaging platforms, the risks of end-to-end encrypted messaging are more targeted. This allows Meta to identify users and rightsholders who face these risks and take targeted actions to mitigate them.

We note that in this section we are covering both contextual factors that may impact the severity and likelihood of risks (e.g., virality), as well as the risks themselves (e.g., human trafficking).

The main categories of risks, each of which comes with a range of human rights harms, are as follows:

**Risk 1: Child Sexual Abuse and Exploitation**

The use of end-to-end encryption across all of Meta's messaging platforms may inhibit the company's ability to detect, remove, and report CSAM, as well as content or accounts related

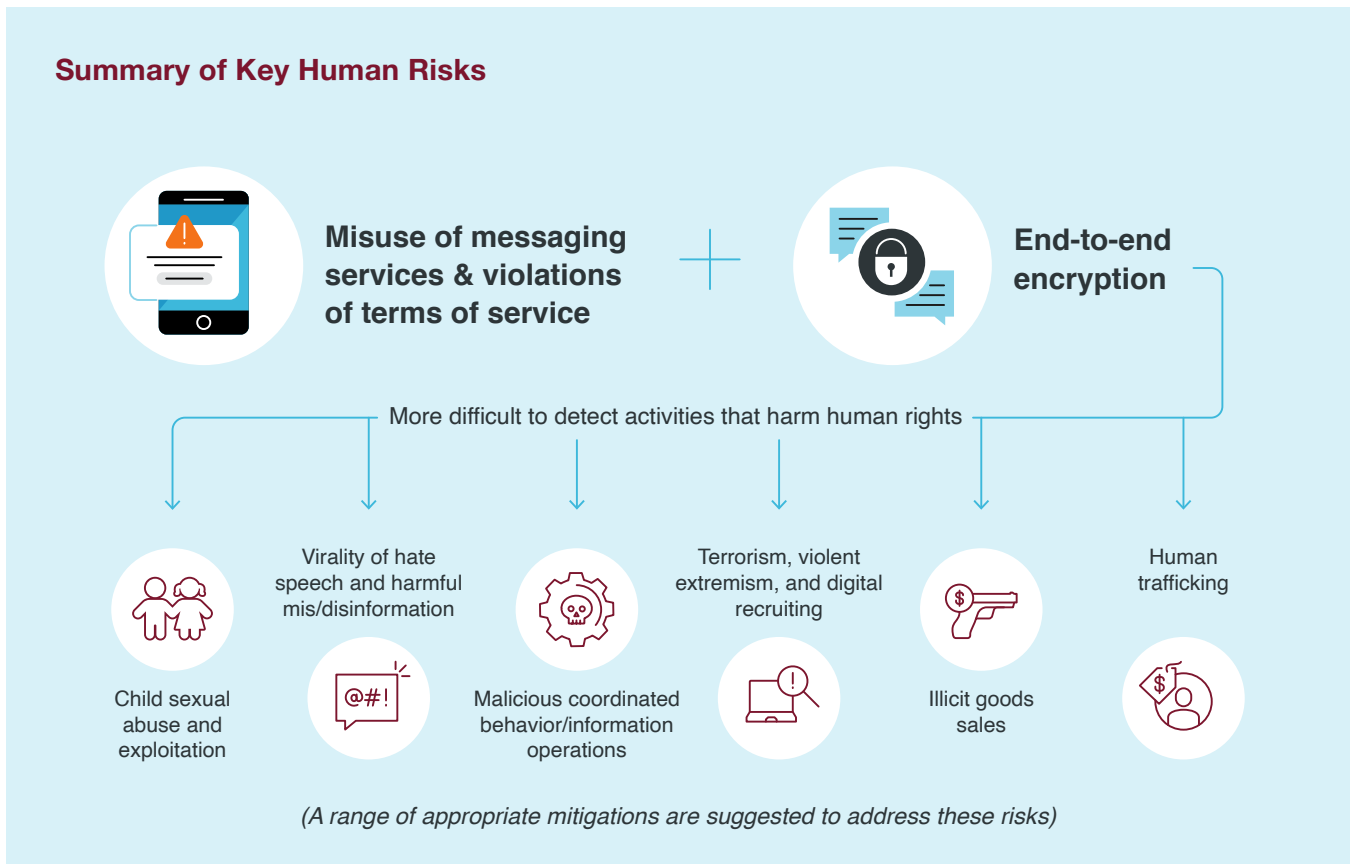
to grooming, sexual extortion of children, child sex tourism, child prostitution, and trafficking of children, among other harms.

**Risk 2: Virality of Hate Speech and Harmful Mis / Disinformation**

While virality in and of itself does not constitute a violation of human rights, it can amplify and spread hate speech and mis / disinformation in a way that leads to, or exacerbates, human rights harm. Viral instances of this content may be more challenging to detect in an end-to-end encrypted environment, therefore making it more difficult to address potential harm.

**Risk 3: Malicious Coordinated Behavior / Information Operations**

Coordinated behavior with malicious intent, both authentic (i.e., by real people using real accounts) and inauthentic (i.e., by people using fake accounts), undermines the integrity of social media platforms and messaging services. While





malicious coordinated behavior is not itself a human rights violation, it can be an enabling factor in bad actor exploitation of messaging services, and can impact rights such as nondiscrimination, bodily security, privacy, freedom of expression, and democratic participation. Malicious coordinated behavior may be more difficult to detect and address in an end-to-end encrypted environment.

- **Risk 4: Illicit Goods Sales**

In addition to legal commerce, a wide range of illicit activity takes place on Facebook Marketplace, in private groups, and via messaging services. These activities can impact bodily security rights, and may be more difficult to detect and address in an end-to-end encrypted environment.

- **Risk 5: Human Trafficking**

End-to-end encrypted messaging may be used to facilitate human trafficking, including but not limited to sex trafficking, labor trafficking, organ trafficking, and child marriage. Constantly switching between different open and closed-communications messaging platforms is a technique that traffickers use to facilitate illegal advertising, recruitment, control, punishment, and coercion of victims.

- **Risk 6: Terrorism, Violent Extremism, and Digital Recruiting**

Violent extremist and terrorist groups have proven to be tech savvy and have used more end-to-end encrypted messaging platforms to communicate with followers, disseminate propaganda, incite violence, and coordinate terrorist attacks that result in loss of life and bodily harm.

## 8.1 Child Sexual Exploitation and Abuse

### Context

Children benefit from the human rights opportunities of end-to-end encrypted messaging, such as increased privacy, greater opportunities for freedom of opinion and expression, and physical safety. However, some of the most severe human rights risks of Meta's expansion of end-to-end encryption involve the use of end-to-end encrypted messaging to facilitate the sexual abuse and exploitation of children. Child sexual abuse and exploitation online consists of multiple overlapping types of harm—specifically, the grooming, sexual exploitation and extortion, and trafficking of children, live streaming of child abuse, self-generated sexual content by children, as well as the undetected distribution of CSAM, which is a violation of the privacy rights and a revictimization of the children involved every time material is shared and viewed.

### Typology of Online Child Sexual Abuse and Exploitation



Citation: [https://ecpat.org/wp-content/uploads/2021/05/SECO-Booklet\\_ebook-1.pdf](https://ecpat.org/wp-content/uploads/2021/05/SECO-Booklet_ebook-1.pdf)

Law enforcement agencies, legislators, child protection organizations, and company shareholders around the world are concerned about what will happen if Meta is unable to readily detect CSAM shared via messaging and take action against accounts used to exploit children once it shifts fully to end-to-end encryption on its messaging services. In October 2019, top law enforcement officials from the US, UK, and Australia sent an open letter to Mark Zuckerberg urging a halt to Meta's plan for implementing end-to-end encryption across all its messaging platforms. They asked Meta to provide "lawful access" to end-to-end encrypted messages to help law enforcement agencies fully protect children online.<sup>1</sup> Some child protection organizations have also called on Meta to delay its expansion of end-to-end encryption until it can continue detecting and reporting CSAM.

Currently, Meta is widely considered to be the most active of any technology company in proactively scanning for and reporting CSAM. Meta accounts for the vast majority of online CSAM reports to the National Center for Missing & Exploited Children (NCMEC), which acts as a clearinghouse to report CSAM detected across the internet to law enforcement in order to save exploited children and prosecute perpetrators. According to NCMEC, in 2018 more than 18.4 million cases of child sexual abuse were reported to their Cyber Tipline, of which 12 million came from Facebook Messenger.<sup>2</sup> NCMEC estimated that more than half of such reports would not be possible if end-to-end encryption were to be implemented.<sup>3</sup>

These absolute numbers are staggering, but also mask nuance about the nature of child sexual abuse material on Meta's platforms. Only a small proportion of the 12 million cases Meta reported to NCMEC were unique—upwards of 90 percent were duplicates of previously reported content—which

means that in the vast majority of cases the same content is being shared over and over again, and the number of individual victims is much smaller than the total amount of CSAM. Despite this, the adverse rights impacts of CSAM should not be minimized—these 12 million cases represent repeated privacy violations and the revictimization of the rightsholders involved. Unfortunately, there is no accurate measure of the scope of child exploitation resulting from CSAM online because it's unclear how many individual victims there are.

Currently, internet companies rely on both content-based and metadata-based technologies to detect and remove CSAM online. One of these content-based technologies is called PhotoDNA,<sup>4</sup> and Meta is one of the main contributors to its Industry Hash Sharing Platform, a hash dataset that enables the use of PhotoDNA and other fingerprinting technologies.<sup>5</sup> By expanding end-to-end encryption, the number of Meta contributions of CSAM to PhotoDNA hash datasets may decrease somewhat. However, because most CSAM detected in Meta's messaging apps is duplicated (i.e., it has previously been detected and added to the Industry Hash Sharing Platform), and because the majority of new CSAM content Meta contributes to the Industry Hash Sharing Platform is detected on its public platforms, Meta's expansion of end-to-end encryption is unlikely to result in a substantial decrease in the amount of new CSAM added to the hash-sharing platform. However, it will certainly result in a decrease of total reports and therefore will likely diminish the ability to identify all users who share CSAM.

It is also important to note that even with end-to-end encryption, Meta can still take action against CSAM unencrypted data associated with messaging. For example, WhatsApp removes hundreds of thousands of accounts per month that

1 <https://www.justice.gov/opa/pr/attorney-general-barr-signs-letter-facebook-us-uk-and-australian-leaders-regarding-use-end>.

2 <https://www.nytimes.com/interactive/2019/09/28/us/child-sex-abuse.html?smtyp=cur&smid=tw-nytimes>.

3 <https://www.missingkids.org/theissues/end-to-end-encryption#bythenumbers>.

4 PhotoDNA is a technology developed in 2009 by Dartmouth College and Microsoft to detect and report CSAM. The technology, which has been adopted and used by major social media companies and NCMEC, creates a digital signature, known as a "hash," for an image. That hash is then compared with the hash datasets of known illegal and previously detected images. If a match is found the platform then decides whether to remove the content and report the issue. See <https://www.microsoft.com/en-us/photodna>.

5 The Industry Hash Sharing Platform is "a cloud-based hash sharing tool, and the first collaborative industry initiative to improve and accelerate the identification, removal, and reporting of child abuse images across different digital networks." <https://www.thorn.org/reporting-child-sexual-abuse-content-shared-hash/>.

are found to be sharing CSAM in their profile or group photos, are reported by users, or are revealed by group metadata.

As with all types of problematic content, the effectiveness of user reporting of child exploitation on Meta's messaging platforms could also be negatively impacted by end-to-end encrypted messaging. Currently, when users report unencrypted messages, Meta is able to see the entirety of the conversation history. However, when users report end-to-end encrypted messages on WhatsApp and Secret Messages, Meta is only able to see recent messages, which may not contain enough relevant content to clearly identify child exploitation. The way in which Meta decides to implement user reporting within its end-to-end encrypted messaging platforms will thus likely have a significant impact.

However, we note that user reporting of child sexual abuse and exploitation is likely to be more effective in detecting inappropriate interactions with children than with CSAM. Meta's research has indicated that a significant portion of CSAM is shared without the intention of exploiting children (e.g., in outrage),<sup>6</sup> and therefore in-app education and user reporting improvement is likely to improve reporting of CSAM in this context. However, CSAM is also often distributed by users who want to receive that kind of content and therefore are unlikely to report it.

Also relevant is Messenger Kids, a messaging platform designed for children under 13 with significant parental controls. Although at the time of writing, Meta has not yet decided whether or not to end-to-end encrypt Messenger Kids, Meta has stated that it is committed to providing parents with the same visibility and control over their child's experience if it does decide to deploy end-to-end encryption in Messenger Kids. This would be done via account linking and providing parents access to a dashboard that enables them to approve new contacts and see any shared media, while preventing access by Meta employees absent a

report. However, this does not ensure that all users under 13 would be protected, as it is common for children to use regular Meta platforms despite the company's age limit.<sup>7</sup>

### Understanding Specific Human Rights Harms Associated with Child Exploitation

Children are rightsholders and have the right to be protected from sexual exploitation and abuse, according to Article 34 of the Convention on the Rights of the Child. International human rights law recognizes that children are an especially vulnerable group, and thus deserving of special consideration. Child rights organizations have urged companies to pay special attention to the human rights of children and put systems in place to ensure children safety, privacy, and other human rights.<sup>8</sup> However, implementing mitigation techniques to protect children in end-to-end encrypted messaging contexts is complex because the exploitation of children takes various forms that result in human rights harms of differing severity.



<sup>6</sup> <https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/>.

<sup>7</sup> <https://www.thorn.org/thorn-research-minors-perspectives-on-disclosing-reporting-and-blocking/>.

<sup>8</sup> <https://www.unicef.org/media/66616/file/Industry-Guidelines-for-Online-ChildProtection.pdf>.

For instance, in the case of circulating known CSAM, the immediate human rights impact is a violation of the privacy and dignity of the child involved because the violations to bodily integrity rights occurred in the past when the CSAM was produced. Nevertheless, the sharing of CSAM over time revictimizes the child and can lead to what are known as cumulative impacts, where one case taken in isolation may not have significant human rights impacts but, when combined with thousands of similar cases, may result in severe human rights impacts. For survivors of child sexual abuse, knowing that their images continue to circulate online results in increased damage to their privacy, can have significant impacts on their dignity and psychological well-being, and can increase their vulnerability to future harm.

The case of users sharing live video of sexual violence against children via end-to-end encrypted video chat constitutes a real-time violation of the rights to bodily integrity and security, and freedom from torture and cruel, inhuman, and degrading treatment. Unfortunately, whereas known CSAM images that have been previously reported and tagged could in theory be detected via tools such as PhotoDNA, detecting live child exploitation in a video chat would be much more technically difficult.

*Assuming Meta does adopt reasonable mitigation measures—such as the recommendations contained in this assessment—then BSR considers Meta to be “directly linked” to (rather than “contributing to”) the potential adverse child sexual abuse and exploitation impacts associated with the expansion of end-to-end encryption.*

As was explored earlier, **assuming Meta does adopt reasonable mitigation measures—such as the recommendations contained in this assessment—then BSR considers Meta to be “directly linked” to (rather than “contributing to”) the potential adverse child sexual abuse and exploitation impacts associated with the expansion of end-to-end encryption.** (see Section 6.1 for more details). Meta is currently very active in detecting and reporting CSAM found on Messenger, and so is aware of the extent of the problem. It is thus reasonably foreseeable that the distribution of CSAM will continue when messaging platforms are fully end-to-end encrypted, and may actually increase as bad actors benefit from the privacy protections of end-to-end encrypted messaging and the ease of finding like-minded people provided by cross-app communication. Absent reasonable mitigation measures (see below), Meta would be responsible for providing or cooperating in remediation for harms, such as by funding rehabilitation groups who work with CSAM victims or through financial compensation for victims.

### Potential Mitigations

#### 1. Using behavioral signals and cross-platform data to prevent exploitation from occurring

Although CSAM detection has been the main focus of many external stakeholders, particularly policymakers and law enforcement actors, there are many other strategies Meta is deploying to prevent other forms of child exploitation in end-to-end encrypted messaging, such as using classifiers and cross-platform information to intervene upstream to prevent violations from occurring in the first place.

One point of intervention is preventing connections between unknown adults and minors from occurring. Using cross-platform data to identify the users' estimated ages, Meta currently seeks to prevent unknown adults from connecting with children by not showing minors in users' “people you may know” feature, and alerting minors when an adult they are connected to may be a risk and suggest blocking and reporting.

Another point of intervention is analyzing user behavior to identify whether the person is behaving like someone in their demographic or whether there are indicators of child exploitation. For example, if a user from one city is contacting a minor in a different city, or if a user spends time looking at the photos of minors, searching for specific keywords, or looking at certain groups, then Meta will remove the “add to friends” button on Facebook and/or hide the minor from search results to prevent the user from ever making contact with the minor.

An additional point of intervention involves analyzing public platform content to identify users seeking to exploit children. For example, users seeking CSAM often use code words in their social media presences to recognize each other, so Meta uses classifiers to search for those code words and then prevents these users from connecting. Another flag is users who post sexualized comments on the public posts of minors.

These signals are often combined and analyzed to identify accounts for further investigation. For example, signing up for an account and then rapidly joining and leaving hundreds of different group messages is a common indicator of accounts that spread CSAM. To verify whether this is likely the case, this information might be cross-referenced with public information such as profile photos and group names, as well as information from user reports, to identify and remove these accounts and groups.

As with any classifier-based system, there is a risk of both false positives and false negatives. For example, false negatives occur when Meta’s classifiers fail to catch inappropriate interactions between adults and children. This is particularly a risk for inappropriate real-world relationships that have transferred to messaging platforms, as is often the case with child sexual abuse and exploitation by family members or friends of the victims. Meta relies on user reports of content to reduce false negatives. False positives occur when users are falsely flagged for a child safety violation and have their accounts suspended or removed. Meta relies on user appeals to reduce false positives.

## 2. Improving user reporting and education

User reporting is an important mitigation for addressing many forms of harmful content in end-to-end encrypted messaging, and child sexual exploitation and abuse is no exception. Making reporting channels more clear and seamless, and improving underage users' awareness, understanding, and trust of reporting can help increase the volume of reports, and thus can help Meta identify instances of unwanted and exploitative interactions with adults. However, it is unlikely to significantly impact the distribution of CSAM, which is largely shared between users who desire and consensually share that type of content and therefore would not report it.

## 3. “Client-side scanning” to detect and block CSAM while maintaining end-to-end encryption

Methods such as client-side scanning of a hash corpus, trained neural networks, and multiparty computation including partial or fully homomorphic encryption (often collectively referred to as “client-side scanning,” although some can also be server-side) have all been suggested as solutions to enable Meta to continue to identify, remove, and report CSAM. However, most proposed methods would undermine the cryptographic integrity and constitute a disproportionate restriction on privacy and other human rights, and therefore should not be pursued. The only method proposed thus far that may not undermine cryptographic integrity (and therefore has the potential to be necessary and proportionate) is homomorphic encryption, which enables the “processing” of data while it is encrypted. However, it is a nascent approach that is far too computationally intensive to implement in a messaging context for even high-end mobile devices today. For example, Meta’s own research of a homomorphic encryption approach found that it would take around 20 million seconds (over seven months) to run. Any technical solution would need to work on low-end devices, which are used by a large percentage of Meta users, to be effective and respect the different circumstances of vulnerable groups.

Research into these methods is still in its nascent stages, and computational power will likely eventually increase enough to enable homomorphic encryption or other solutions that maintain cryptographic integrity. Meta and many other stakeholders argue that any form of CSAM scanning—even that which doesn't undermine cryptographic integrity—is inconsistent with the principles of end-to-end encryption. They also point out that even a method of client-side scanning that maintains cryptographic integrity could be abused by bad actors and lead to other human rights risks.

For example, if Meta starts detecting and reporting universally illegal content like CSAM, some governments are likely to exploit this capability by requiring Meta to block and report legitimate content they find objectionable, thereby infringing on the privacy and freedom of expression rights of users. It is noteworthy that even some prior proponents of homomorphic encryption have subsequently altered their perspective for this reason, concluding that their proposals would be too easily repurposed for surveillance and censorship.<sup>9</sup> In addition, these solutions are not foolproof; matching errors can occur, and bad actors may take advantage of the technical vulnerabilities of these solutions to circumvent or game the system. It is therefore BSR's recommendations that if the implementation of client-side scanning solely to detect CSAM—a legitimate aim—would likely result in a significant restriction of freedom of expression, privacy, and other rights, then client-side scanning should not be pursued. This is discussed in more detail in Sections 10 and 11.

#### **4. Engaging with external stakeholders and supporting online child safety efforts**

In addition to taking action on its platforms, Meta also engages with NCMEC and other external partners to better understand the nature of child exploitation online and share technical knowledge

about detection and interruption. This includes working with NCMEC to help triage the huge amount of CSAM reports and identify instances of new content or indicators of active abuse so that authorities can act; voluntarily collaborating with law enforcement entities to improve their capability to identify and investigate cases of child sexual abuse and exploitation in rights-respecting ways; working with other platform companies to develop common approaches to the internet-wide challenge of child safety; and conducting collaborative, transparent research to better understand the problem.<sup>10</sup>

#### **IMPACT FACTORS**

**Age plays an important role in the vulnerability of children to exploitation. Children between the ages of 13 and 18 who can officially become users of Meta products might be more in danger of specific types of exploitation such as grooming, sexual extortion, and trafficking than children under the age of 13 whose parents can better protect them on messaging services by using programs such as Messenger Kids. However, it is important to note that many children under 13 still use regular messaging products, despite Meta's age limit, due to the notorious difficulty of verifying users' age faced by all online platforms.**

**Laws governing child exploitation, as well as the capacity of law enforcement, differ greatly between countries and regions. The scale and scope of negative human rights impacts is affected by the type of exploitation, age, gender, level of digital literacy, and law enforcement competency in different countries. Currently, over 90 percent of Meta's CSAM reports to NCMEC originate from countries outside of the US. Although these reports are shared with national points of contact around the world to act upon, not all countries have the same capacity to take effective action. Child exploitation is a global, system-wide problem that companies like Meta can only address so much in the absence of capable government authorities.**

<sup>9</sup> See <https://www.washingtonpost.com/opinions/2021/08/19/apple-csam-abuse-encryption-security-privacy-dangerous/>.

<sup>10</sup> Facebook already undertakes significant research on its own, such as <https://research.fb.com/blog/2021/02/understanding-the-intentions-of-child-sexual-abuse-material-csam-sharers/>.

## 8.2 Virality of Hate Speech and Harmful Mis / Disinformation

Virality is an innate element of social media platforms where everything from practical hacks to inspiring stories can reach millions of users in a matter of hours. While virality itself is neutral, it can amplify and spread problematic content in a way that leads to or exacerbates harm. This kind of content can lead to human rights violations, especially if it involves hate speech, harmful misinformation or disinformation, or incitement of violence that leads to offline harm. Conflict-affected or high-risk areas, as well as countries with fragile information ecosystems, are most at risk of viral content leading to offline harm.

It should also be noted that content not violating Meta's Community Standards can still result in adverse human rights impacts. For example, one case of "borderline" content, taken in isolation, may not have significant human rights impacts but, when combined with thousands of similar cases, may result in severe human rights impacts.

Viral hate speech has some specific additional human rights risks. It can affect freedom of thought and religion if the hate speech encourages users to harass and harm members of a religious group in a way that prevents them from openly gathering and practicing their religion without fear of violence. Viral hate speech can also impact free association and assembly, the right to participate in government, and the right to freely participate in cultural life if people targeted by hate speech feel they cannot safely gather with others, vote or participate in political events, or freely participate in cultural life. It can also impact free expression if users in group messages feel they must self-censor for fear of retribution.

When considering the potential human rights impacts of viral content, it is important to understand its severity and likelihood. An analysis of harmful content on Messenger found that high severity content (e.g., CSAM) has a relatively low prevalence on the service while lower severity content (e.g., certain types of misinformation) has

a relatively high prevalence. However, even low severity content can have a significant impact in aggregate—for example, one rumor or one misinformation post itself is not as severe as a single CSAM image, but the total adverse impact on society can be significant if the post goes viral.

With the expansion of end-to-end encryption, Meta's limited visibility will make it more challenging to proactively detect and respond to harmful viral content spread via messaging. WhatsApp and Messenger have addressed this issue by limiting the number of times a message can be forwarded to just five times, and WhatsApp has seen a 70 percent reduction in the spread of viral content as a result.<sup>11</sup> WhatsApp also uses other signals of abuse, such as "spammy" behavior or high levels of user reports, as well as information from other Meta platforms, to identify and flag potential bad actor accounts.

In addition to detection and enforcement, WhatsApp has taken several steps to address contextual risks of virality. When a piece of content is highly forwarded, that content is labeled to encourage users to verify the information by showing a magnifying glass icon that directs the user to a pre-populated Google search when clicked on. To address specific virality risks in certain countries, WhatsApp has also built partnerships with fact checkers in countries where it is heavily used. In these markets, users can send content to fact checkers and get a response, as well as access a database of existing fact checks via a chatbot. WhatsApp also runs classifiers on user reports that are related to civic issues to identify trends that might have implications for election integrity around the world.

Additionally, as with all types of problematic content, the effectiveness of user reporting of harmful viral content such as hate speech or incitement to violence on Meta's messaging platforms could also be negatively impacted by end-to-end encrypted messaging. Currently, when users report unencrypted messages on Facebook

11 See <https://www.theverge.com/2020/4/27/21238082/whatsapp-forward-message-limits-viral-misinformation-decline>.

Messenger and Instagram DMs, Meta is able to see the entirety of the conversation history. However, when users report end-to-end encrypted messages on WhatsApp and Secret Messages, Meta is only able to see recent messages, which may not contain enough relevant content to clearly identify problematic content. The way in which Meta decides to implement user reporting within its end-to-end encrypted messaging platforms will thus likely have a significant impact.

## IMPACT FACTORS

**Different regions will have different norms around sharing information (forwarding messages, images, links, size of networks, types of groups). It will be important for Meta to consider geographic context and vulnerability when developing approaches to minimize virality.**

**Access to news and information varies across regions, even down to the local level. Limited access to news outside of Meta's products may increase usage of the products in higher risk ways.**

**Countries with fragile information ecosystems are most at risk of viral content leading to offline harm because the content cannot be countered by authoritative sources. In addition, conflict-affected and high-risk areas are similarly vulnerable to viral content inciting violence.**

### 8.3 Malicious Coordinated Behavior / Information Operations

Certain forms of coordinated behavior, both authentic (i.e., by real users using legitimate accounts) and inauthentic (i.e., by users using fake accounts), can undermine the integrity of social media platforms and messaging services when they mislead others about who posters are or what they are doing. While much of this behavior is driven by financial motivations, some actors seek to covertly influence the opinions, beliefs, and actions of other users (i.e., information operations). The use of fake accounts, shadow accounts, black PR firms,<sup>12</sup> and

other tactics to mislead people is already against Meta's Community Standards. Although Meta is active in detecting information operations on its *public platforms*, end-to-end encryption may make it difficult for Meta to identify such behavior in its *messaging services*.

While coordinated behavior is not in itself a human rights violation (it can be legitimately used as part of activism and public service campaigns, for example), it can be an enabling factor in malicious actors' exploitation of information networks to carry out human rights violations. Malicious coordinated behavior is often designed and executed with the intention of misleading a population, and this interferes with the freedom of opinion and the right to seek, receive, and share information. Malicious coordinated behavior can be used to infringe on the right to free expression by silencing people or coercing them into self-censorship, and can also be used to invade the right to privacy by harassing and doxing individuals. In addition, some forms of malicious coordinated behavior may be used to violate bodily security rights by inciting violence or hostility against certain groups, often those who are vulnerable.

In contexts with fragile information ecosystems, malicious coordinated behavior can also amplify existing tensions, conflict, and mistrust in the community. For example, Facebook pages that were seemingly independent news or opinion sources were used to covertly push the messages of the Myanmar military. This type of coordinated behavior on social media has been used as a tool by actors who perpetrated genocide and ethnic violence to cover up, downplay, or refute the truth of their actions.<sup>13</sup>

Meta's ability to detect malicious coordinated behavior will not be significantly affected by end-to-end encryption because the majority of coordinated behavior currently occurs on Meta's public platforms. However, Meta's overall visibility will be limited because Meta will not be able to see

<sup>12</sup> <https://www.buzzfeednews.com/article/craigsilverman/disinformation-for-hire-black-pr-firms>.

<sup>13</sup> See BSR's HRIA of Facebook in Myanmar as an example: [https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria\\_final.pdf](https://about.fb.com/wp-content/uploads/2018/11/bsr-facebook-myanmar-hria_final.pdf).



what content such coordinated actors are pushing in end-to-end encrypted messages. This may limit investigatory opportunities to understand the narratives pushed by malicious coordinated actors, the evolution of their tactics, or how it may align with other coordinated behavior across Meta or other platforms.

Despite this challenge, Meta is able to address malicious coordinated behavior in end-to-end encrypted messaging by connecting actors' cross-platform and behavioral signals. For example, if Meta detects accounts engaging in coordinated inauthentic behavior on the Facebook platform (which is a violation of the Facebook Community Standards) and those accounts have phone numbers tied to WhatsApp accounts, then that information will be shared with WhatsApp. WhatsApp will then review the available evidence, and if it determines that WhatsApp is being used for coordinated inauthentic behavior in a way that is leading to harm, WhatsApp will ban the accounts.

## IMPACT FACTORS

**There are no specific vulnerable groups impacted by malicious coordinated behavior, as this category encompasses a wide variety of content. However, multiple groups (e.g., children, religious or ethnic minorities, human rights defenders, political opponents, or journalists) could be at greater risk, depending on the types of content and the target audiences of the actors.**

**In some regions, government entities may be the perpetrator of malicious coordinated behavior, either domestically or in other countries. For example, based on takedown reports from Meta, countries such as Egypt, Iran, and Russia have engaged in multiple coordinated campaigns targeting other countries.**

### 8.4 Illicit Goods Sales

Facebook was originally created as a social network, not a marketplace for e-commerce—yet, as Meta products have become more ubiquitous around the world, a wide range of online activity has

moved onto the platforms. In addition, Meta now offers Facebook Marketplace and has several other features that support commercial pages and the sale of goods, including the ability to store users' credit card information, the ability to send peer-to-peer payments through Messenger and Instagram, the integration of Buy and Shop buttons with ads and messages, and the growth of the WhatsApp Business API.

In addition to legal commerce, public reporting suggests that illicit activity also sometimes takes place on Facebook Marketplace, in private groups, and via messaging services, despite the fact that such activities violate Facebook's terms of service and often local laws. Sales on Facebook and Instagram are incredibly decentralized, making it difficult to monitor and track illegal goods such as weapons, drugs, exotic animals, human organs, and cyber-fraud services. Meta has developed a range of detection and prevention focused mitigations to address illegal goods sales, but cannot catch them all. These activities can have human rights impacts when they involve crimes that violate the personal security or bodily integrity of victims.

These exchanges are often undertaken by individuals who seek out ways to remain anonymous and hide their interactions and activities. BSR anticipates that unless properly mitigated, the move to end-to-end encryption will increase the opportunities for bad actors to carry out these transactions with impunity as users conducting criminal activity benefit from the privacy protections of end-to-end encryption and the ease of connecting with others facilitated by cross-app communication. As with other types of problematic content mentioned previously, Meta can use behavioral signals and connect violating accounts across platforms in order to identify accounts involved in illegal goods sales using end-to-end encrypted messaging. Meta is currently examining various mitigations to address the risk in Marketplace-connected messages on Facebook Messenger, such as including Meta as a party in each conversation connected to a Marketplace ad and clearly disclosing this to users. At the time of

writing, there is no end-to-end encryption as part of Facebook Marketplace.

Additionally, as with all types of problematic content, the effectiveness of user reporting of illicit goods sales on Meta's messaging platforms could also be negatively impacted by end-to-end encryption. Currently, when users report unencrypted messages, Meta is able to see the entirety of the conversation history. However, when users report end-to-end encrypted messages on WhatsApp and Secret Messages, Meta is only able to see recent messages, which may not contain enough relevant content to clearly identify illicit goods sales. The way in which Meta decides to implement user reporting within its end-to-end encrypted messaging platforms will thus likely have a significant impact.

### IMPACT FACTORS

**There are no specific vulnerable groups implicated by illicit goods sales, but multiple groups (e.g., children, the elderly, etc.) could be at greater risk if they are taken advantage of by malicious sellers, or if the purchased goods lead to some secondary harm. In some regions, Meta platforms may be a more accessible way to buy and sell illegal goods compared to other online markets and places, and may therefore deal with a higher amount of illicit goods sales. Countries with weak rule of law are more likely to have robust sales of illegal goods both online and offline. In addition, various e-business features of each Meta messaging platform could complicate efforts to understand how each platform is used in illicit goods sales between parties who are sometimes in different regions of the world.**

## 8.5 Human Trafficking

Human trafficking is among the crimes that could become more prevalent and difficult to detect on end-to-end encrypted messaging platforms. At the height of refugee migration to Europe in 2017,

the UN's International Organization for Migration (IOM) warned that platforms such as Facebook Messenger and WhatsApp have created a path for human traffickers to lure West African migrants, and Syrian refugees in the Middle East have been sold as slaves via WhatsApp.<sup>14</sup> This potentially puts people at risk of the various human rights harms associated with trafficking, including detention, sexual abuse, torture, slavery, and even death.<sup>15</sup>

It is important to note that end-to-end encrypted messaging can also benefit refugees fleeing their countries as well as victims of trafficking. According to anti-trafficking organizations, victims regularly use Meta's messaging platforms to connect with service providers and find their way out of trafficking. The privacy protections of end-to-end encryption ensures they are able to securely connect with service providers. However, in this section we focus specifically on human rights risks.

Traffickers often switch between different unencrypted and encrypted messaging platforms to facilitate illegal advertising and recruitment, and to control, punish, and coerce victims. Currently, while Meta seeks to proactively detect trafficking on its public platforms, which can lead to action taken on messaging, Meta does not proactively detect instances of trafficking on its messaging platforms alone. This means that any trafficking caught in messaging is the result of reporting by users or by anti-trafficking groups that are part of the Trusted Partners program and report trafficking content, or via information gleaned from public parts of the platform. This means that, unlike with CSAM, Meta is not as aware of the prevalence of trafficking on its messaging platforms. However, according to a survey run by the Polaris Project, messaging apps broadly were one of the most commonly reported online platforms used for facilitating the recruitment of victims into trafficking and modern slavery.<sup>16</sup>

Unlike in the exploitation of children, for which

<sup>14</sup> <https://ourworld.unu.edu/en/women-are-being-traded-as-slaves-on-whatsapp-heres-how-the-un-can-act>.

<sup>15</sup> <https://www.dw.com/en/un-migration-agency-urges-facebook-to-combat-human-traffickers/a-41716014>; <https://www.unodc.org/e4j/en/tip-and-som/module-3/key-issues/investigative-and-prosecutorial-multidisciplinary-approaches.html>.

<sup>16</sup> <https://polarisproject.org/resources/on-ramps-intersections-and-exit-routes-a-roadmap-for-systems-and-industries-to-prevent-and-disrupt-human-trafficking/>.

Meta has more clear policies,<sup>17</sup> Meta's approach to combating human trafficking is less definitive. This is because trafficking is more complex and ambiguous than child exploitation, and lacks an equivalently strong legal structure. Trafficking takes different forms across industries, from labor trafficking for tourism and domestic work and trafficking of refugees and migrants to sex trafficking. Sex trafficking can be particularly difficult to detect because of challenges identifying the difference between the consensual exchange of sexual content vs. sexual content shared in the context of trafficking. All of these issues make definition detection of trafficking difficult even on public platforms, let alone in end-to-end encrypted messaging. The complexity of detecting trafficking will make it difficult for Meta to rely solely on "behavioral signals" to detect and act upon accounts related to trafficking, which reduces the likelihood that Meta will be able to identify trafficking in a fully end-to-end encrypted environment through behavioral means alone.

Additionally, the effectiveness of user reporting of trafficking on Meta's messaging platforms could be negatively impacted by end-to-end encryption. Currently, when users report unencrypted messages, Meta is able to see the entirety of the conversation history. However when users report end-to-end encrypted messages on WhatsApp and Secret Messages, Meta is only able to see recent messages, which may not contain enough relevant content to clearly identify trafficking. The way in which Meta decides to implement user reporting within its end-to-end encrypted messaging platforms will thus likely have a significant impact.

## IMPACT FACTORS

**Trafficking especially affects the rights to safety and security of vulnerable groups that are protected under various international human rights conventions, including women, refugees and migrants, LGBTQIA+ communities, and minors. Other socioeconomic factors such as poverty, gender, age and the level of literacy (including digital literacy) play a role as well because navigating user reporting sometimes requires a higher level of digital literacy. In addition, trafficking is also a higher risk for people who live in contexts of extreme poverty and conflict.**

### 8.6 Terrorism, Violent Extremism, and Hate Groups

In recent years, terrorist and violent extremist groups have begun to use end-to-end encrypted platforms. This shift occurred partially in response to decisions by social media companies to ban extremist groups from their platforms and US legal requirements that companies remove certain sanctioned terrorist groups, as well as the increasing prevalence and utility of end-to-end encrypted messaging apps in general.<sup>18</sup> Terrorist and violent extremist groups and individuals have since used end-to-end encrypted messaging platforms to communicate with followers, disseminate propaganda, incite violence, and coordinate terrorist attacks that result in loss of life and bodily harm. Meta still finds terrorist content on its unencrypted messaging platforms, but it is a small minority of the total terrorist content found across Meta platforms.

Similar to child exploitation issues, law enforcement agencies and legislators have urged companies to provide backdoor access to their end-to-end encrypted platforms in order to prosecute terrorist crimes.<sup>19</sup> These officials believe that access to message content is necessary. For example, in his testimony to the US Congress in 2017, FBI

<sup>17</sup> Facebook's policies on child exploitation:

Instagram policy: <https://help.instagram.com/423234141135444>.

Facebook policy: [https://www.facebook.com/communitystandards/child\\_nudity\\_sexual\\_exploitation](https://www.facebook.com/communitystandards/child_nudity_sexual_exploitation).

WhatsApp policy: <https://faq.whatsapp.com/en/165022051727702/>.

<sup>18</sup> <https://www.techagainstterrorism.org/2021/09/07/terrorist-use-of-e2ee-state-of-play-misconceptions-and-mitigation-strategies/>.

<sup>19</sup> <https://money.cnn.com/2017/06/04/technology/theresa-may-london-attack-internet/index.html>.

director Christopher Wray noted that “non-content information, such as metadata, is often simply not sufficient to meet the rigorous constitutional burden to prove crimes beyond a reasonable doubt. Developing alternative technical methods is typically a time-consuming, expensive, and uncertain process.”<sup>20</sup> In the absence of access to message content, it is likely that law enforcement will begin requesting metadata on a wider range of accounts, which will bring new risks to the privacy rights of users that Meta will need to mitigate.

Many compare terrorism and violent extremist content with CSAM, but they differ in three key ways: (1) CSAM is relatively clear and easy to define, whereas terrorist content is much more nebulous, nuanced, and contextual; (2) whereas it is never acceptable and usually illegal to share CSAM, terrorist content can be shared legitimately, for example in journalistic reporting, for counter-speech, or in condemnation; (3) those sharing CSAM typically want to remain hidden, while those sharing terrorist content are often seeking attention.

Taken in combination, these factors make it more challenging for Meta to identify and remove terrorist and violent extremist content and accounts in end-to-end encrypted messaging. For example, the second of these factors means that not every user who shares terrorist or violent extremist content is a member of a terrorist group or is at risk of causing offline harm.

Because end-to-end encryption will prevent Meta from identifying terrorist content in messaging, Meta is focusing on strategic network disruption by identifying and disrupting terrorist and violent extremist users and networks. To do this, Meta is investing in classifiers that use behavioral signals, public platform information matching, and open source intelligence from outside of Meta to identify potential users and groups. For example, it may flag a user because indicators on their public profiles suggested they are at risk of committing a mass casualty attack, because the user was part

of group messages that were removed for terrorist content based on their group photo or description, or because open source intelligence indicated the user was active in extremist groups in other online spaces.

However, this approach will be more successful for certain kinds of terrorist and violent extremist actors than for others, and the trends of terrorism and violent extremism online have evolved in recent years.

The first organizations Meta enforced against were Islamist extremist organizations whose members sought a public profile to reach a wide audience. They use Messenger and WhatsApp, but also frequently have public profiles and content that make them easier to identify and take action against. Because of this, Meta’s expansion of end-to-end encryption will likely not have as significant an impact on identifying and removing users from these groups.

By contrast, more recently right-wing extremism has become a growing threat, and experts both inside and outside of Meta have found that members of these groups tend to behave differently. They are typically more heavily reliant on messaging because their primary audience is their “in-group.” These users rarely utilize public profiles, and when they do, they tend to use coded language that is difficult to detect. Meta’s expansion of end-to-end encryption is therefore more likely to negatively impact its ability to identify and remove right-wing extremism and terrorism users.

Additionally, an overreliance on behavioral signals, coupled with legal requirements to remove terrorist accounts or content from specific sanctioned groups, can lead to over-enforcement that disproportionately affects Muslim communities, and results in over-policing of Islamist extremist content and under-policing of right-wing and other types of terrorist content. Behavioral signals for identifying terrorist and violent extremist accounts are not highly accurate, and without signals from a public

<sup>20</sup> <https://www.fbi.gov/news/testimony/keeping-america-secure-in-the-new-age-of-terror>.

profile or other sources, additional human review may not provide clarity about whether a user is a member of a terrorist group or not.

However, US and European legal requirements incentivize Meta to remove the account even if it does not have a high confidence level. This means some users will be removed from platforms in error, and because the legally sanctioned terrorist groups are largely Islamist extremist organizations, these errors will disproportionately impact Muslim users who are falsely flagged. In many contexts organizations designated as terrorist organizations are also prominent political actors, such as Hamas in Gaza, and there is a lot of legitimate discussion about them online. Because classifiers often involve connecting public platform activity with messaging metadata, this type of legitimate discussion can be swept up as well.

The risks of end-to-end encrypted messaging could be compounded by cross-app communication, which may facilitate terrorist and hate group recruitment by making it easier for members of those groups to identify and connect with others across platforms. At the same time, however, cross-app communication will also make it easier for Meta to use behavioral signals to enforce against terrorism content across platforms, making cross-app communication a double-edged sword. For

example, a user found sharing extremist content on the Facebook platform could have their WhatsApp account suspended or restricted.

Finally, as with all types of problematic content, the effectiveness of user reporting of terrorism content on Meta's messaging platforms could also be negatively impacted by end-to-end encryption. Currently, when users report unencrypted messages, Meta is able to see the entirety of the conversation history. However, when users report end-to-end encrypted messages on WhatsApp and Secret Messages, Meta is only able to see recent messages, which may not contain enough relevant content to clearly identify terrorist and violent extremist content. The way in which Meta decides to implement user reporting within its end-to-end encrypted messaging platforms will thus likely have a significant impact.

#### IMPACT FACTORS

**Members of vulnerable groups, such as religious or ethnic minorities, are more likely to be the targets of both terrorist attacks and hate crimes. This is likely to be exacerbated in geographic contexts with a long history of sectarian conflict. Additionally, users who live in countries where terrorist groups are active are more likely to be swept up in over-enforcement of terrorism content rules.**

# Personas and Scenarios

The UN Guiding Principles on Business and Human Rights expect companies to pay particular attention to the rights and needs and challenges faced by individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized.

Meta's expansion of end-to-end encryption will be different for different users and rightsholders in different contexts. A thorough analysis of the varying geographic and group-based impacts of end-to-end encrypted messaging is beyond

*Vulnerability is heavily dependent on context, and this means that the specific human rights impacts of Meta's expansion of end-to-end encryption will be different for different users and rightsholders in different contexts.*



the scope of this report. However, here we use hypothetical personas and scenarios to highlight how the expansion of end-to-end encryption, as well as the decisions Meta makes about the expansion of end-to-end encryption, could disproportionately impact the rights of vulnerable groups in different contexts, both positively and negatively. Note that these scenarios are not comprehensive, nor are all vulnerable groups covered.

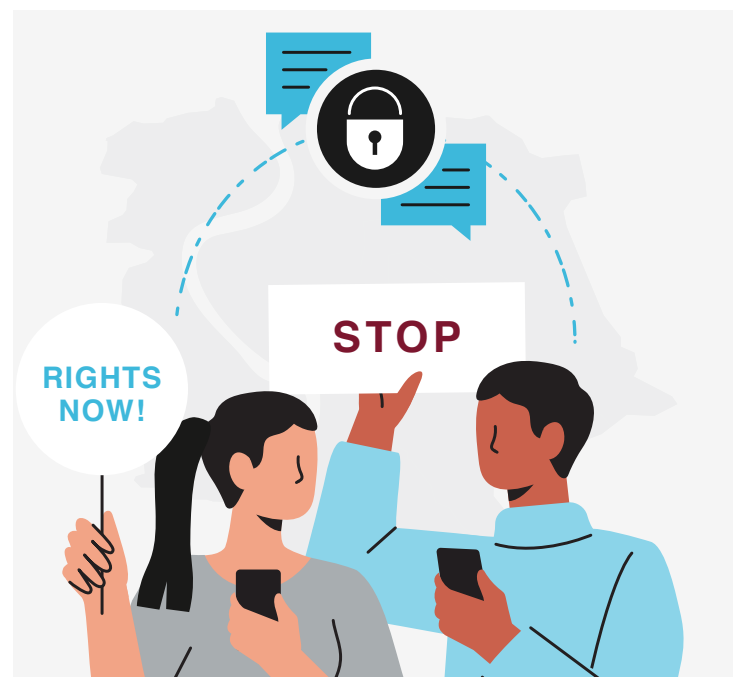
Meta Research started a program called Web-Enabled Simulation to simulate and analyze user behavior in response to different scenarios, and BSR hopes that this section of the HRIA can assist Meta Research in the development of scenarios that prioritize the human rights of vulnerable groups.<sup>1</sup> Although names, locations, and events are invented, these scenarios are all informed by real world circumstances.

<sup>1</sup> <https://research.fb.com/wp-content/uploads/2020/04/WES-Agent-based-User-Interaction-Simulation-on-Real-Infrastructure.pdf>.

### Scenario 1 (Human Rights Opportunity):

## Human rights defenders in an authoritarian country organizing and sharing information via end-to-end encrypted messaging

The government in Country X does not tolerate dissent. Human rights defenders are regularly arrested for speaking out against the government, so activists often choose to use end-to-end encrypted messaging apps such as Signal and Telegram to safely coordinate and organize. They know the government regularly monitors SMS messages and phone calls and so they do not feel safe doing their work on unencrypted channels. However, WhatsApp is the messaging app of choice for the majority of people living their day-to-day lives. Prior to the deployment of end-to-end encryption on WhatsApp, these activists had to use unencrypted channels to conduct much of their daily communications and meaningfully participate in society, making them constantly at risk of being exposed. When end-to-end encryption was deployed on WhatsApp, activists in Country X were not only able to conduct their daily communications securely, but they were also able to safely interact with and reach larger segments of the population who were already on WhatsApp. This was a huge boon to their work and ultimately helped their activism reach a larger audience than it ever had before.



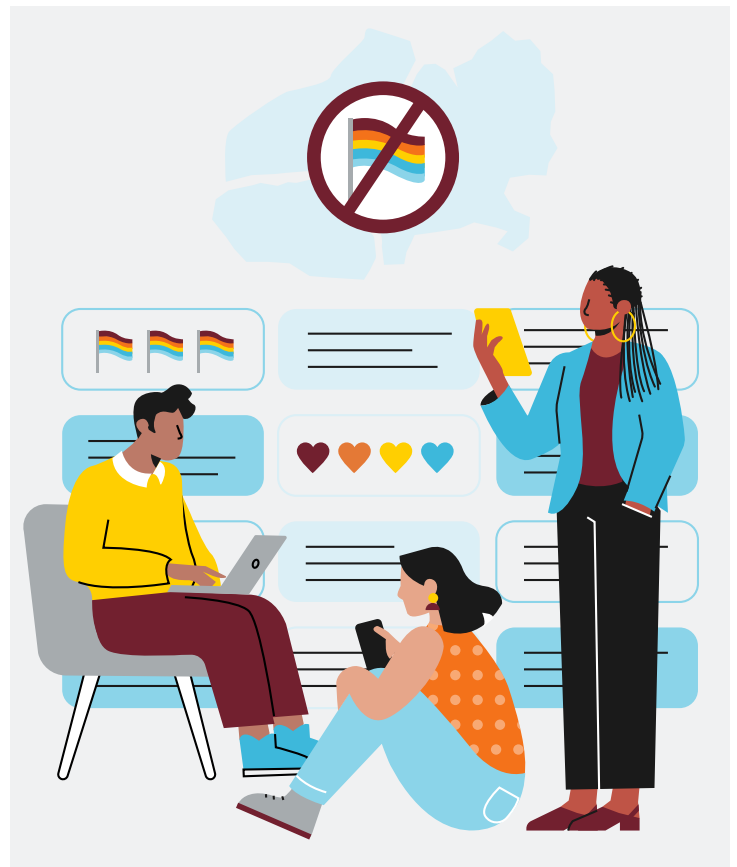
The above scenario illustrates how important ubiquitous end-to-end encryption is for human rights defenders. Because people often use various messaging platforms to communicate, leaving commonly used channels such as Messenger and Instagram DMs unencrypted is a vulnerability for human rights defenders, who increasingly rely on messaging platforms to organize and conduct their work. Although the percentage of Meta users who are human rights defenders is small, as a vulnerable group they require special care and protection.

The scenario also highlights the centrality of the right to privacy in fulfilling other rights, such as freedom of assembly and association, freedom of expression, participation in government, and the right to safety and security. Other vulnerable groups are similarly dependent on the right to privacy to enable these other rights. They include investigative journalists, marginalized racial, ethnic, and religious groups, individuals in abusive relationships and victims of trafficking who use messaging platforms to seek help, and civil society organizations, particularly those focused on women and LGBTQIA+ rights groups, among others.

### Scenario 2 (Human Rights Opportunity):

## Enabling free expression and access to information while protecting physical safety in a repressive environment

In Country Z, homosexuality is illegal and not socially accepted. Members of the LGBTQIA+ community are regularly subject to violent attacks and sometimes killed, and even speaking about homosexuality can lead to imprisonment. Because of this, the community often relies on digital means to connect, gather, and share information. However, prior to end-to-end encrypted messaging, LGBTQIA+ activists and community educators were often identified and targeted via SMS messages, which the government was able to easily intercept due to its ownership over the major telecommunications company. Since having easy access to end-to-end encrypted messaging, the LGBTQIA+ community has been able to connect with each other and share information without fear of reprisals.



This scenario highlights how end-to-end encryption can be especially vital for members of traditionally marginalized and repressed groups, such as the LGBTQIA+ community, and how the privacy protections provided by end-to-end encrypted messaging can also protect people from physical harm. It also highlights how end-to-end encrypted messaging can enable free expression, access to information, and freedom of association in environments where those rights are restricted.



### Scenario 3 (Human Rights Risk):

## The spread of hate speech across multiple platforms

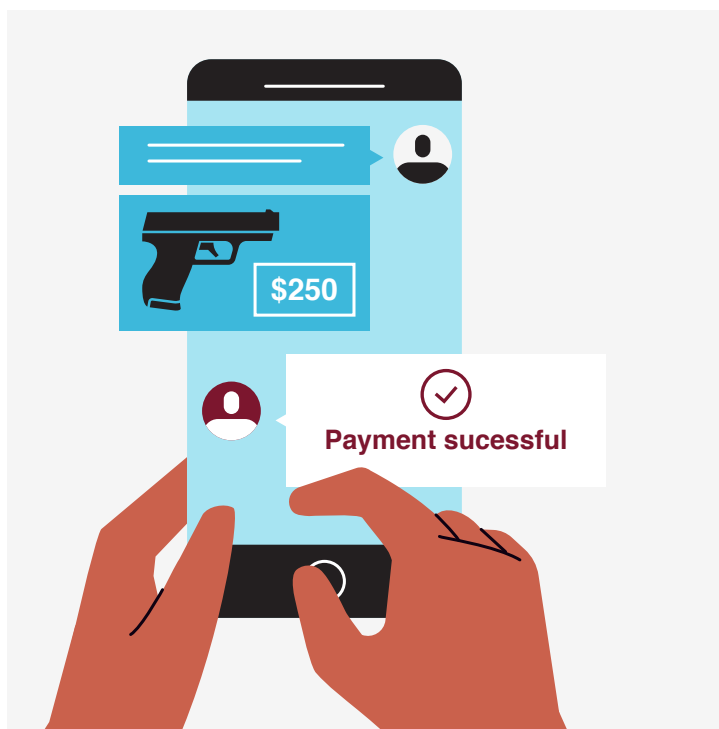
Hateful content denigrating a minority religious group is spread across the Facebook platform after starting in a closed group on Facebook. When the content begins to be taken down for violating the Community Standards of Facebook pages and groups, it spills over onto Messenger, which is now end-to-end encrypted. Hateful messages and the photos, addresses, and other personal information of individuals are shared rapidly due to the ease of cross-app communication via Messenger, WhatsApp, and Instagram DMs. The rapid spread of hate speech contributes to rising tensions and increased division, culminating in riots in which multiple people are severely injured and some are killed.



The above scenario highlights the risk of virality of hate speech and doxing incidents when they move from one platform to another in a very short amount of time. Recent research has just begun to reveal the complexities of addressing hateful content in a timely manner when it crosses platforms within and between different end-to-end encrypted and unencrypted messaging services. Similar scenarios that ultimately culminate in bodily harm can also occur with the viral spread of other types of problematic content. For instance, hoaxes and misinformation can threaten people's health (e.g., misinformation surrounding COVID-19), foreign actors can engage in coordinated inauthentic behavior aiming to foment conflict between different minority groups, or child sexual abuse material can be circulated.

**Scenario 4 (Human Rights Risk):****Illicit goods sales on messaging platforms with different features**

An individual finds an Instagram user who is advertising the sale of firearms. The individual initiates a private message with one of the group members on the now end-to-end encrypted Instagram DMs, where the two share photos, prices, and addresses. To solidify the final step of purchasing the firearm, the two parties move to Facebook Messenger, to utilize its in-app payment feature. The individual purchases a firearm and ultimately uses it to commit a mass shooting.

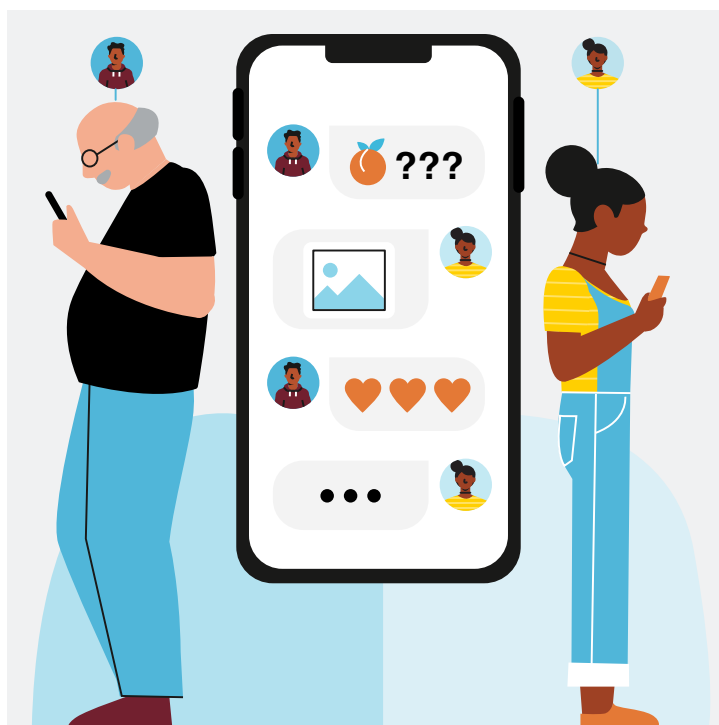


This scenario illustrates how the various business features of each Meta messaging platform (including in-app payment, e-commerce, etc.) could complicate efforts to understand how each platform is used in illicit goods sales between parties who are sometimes in different regions of the world.

## Scenario 5 (Human Rights Risk):

## Undetected grooming and sexual exploitation of a minor

An adult man has been posting increasingly suggestive comments on the Instagram profile of a teenage girl under a fake account posing as a teenage boy. He eventually DM's her on the newly encrypted Instagram direct. He builds her trust slowly over time, eventually requesting she send him nude photos of herself. She complies, regularly sending him photos under the impression they are starting a relationship. All the while, she is unaware of his age or his intentions. Because Instagram is end-to-end encrypted and the man was using a fake account, Meta's behavioral signals tracking did not pick up and prevent the interaction from occurring. Even if Meta had been using a client-side solution to detect child sexual abuse material, the interaction would not have been caught because the photos the girl shared constituted new rather than known and previously reported CSAM, and so would not have been in a hash database.



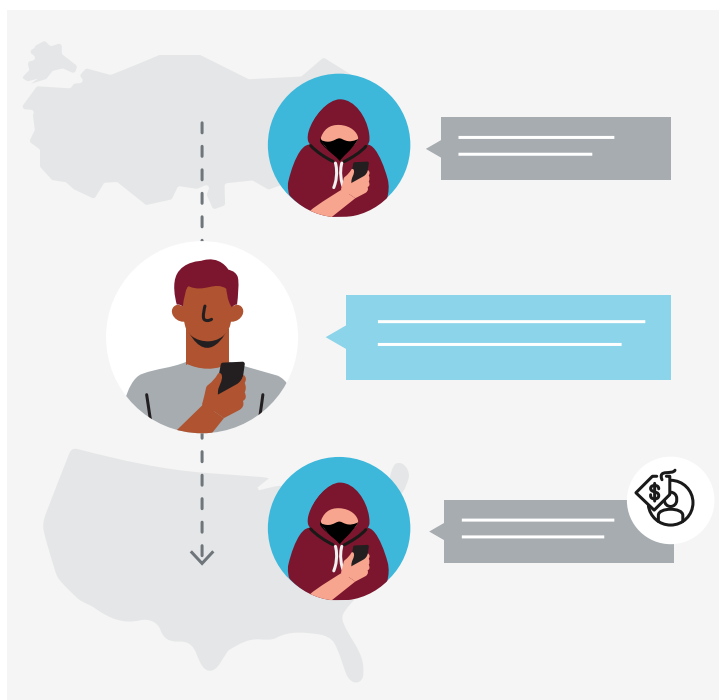
Although end-to-end encryption can create additional challenges to detecting certain kinds of child exploitation, such as the sharing of known CSAM, this scenario illustrates some of the challenges in addressing child sexual exploitation online in general, regardless of end-to-end encryption. In this case, there was nothing to indicate to Meta that the interaction was an exploitative one, and it likely would not have been caught even in an unencrypted messaging service. This scenario also points to the importance of user education and online safety initiatives designed for different age groups.

Scenarios 6 and 7 shed light not on the direct impact of end-to-end encrypted messaging, but on some of the potential human rights impacts that might arise as a result of cross-app communication of the messaging platforms and the ways in which Meta can detect problematic content in an end-to-end encrypted environment. These hypothetical scenarios highlight how overreliance on machine learning-enabled systems and user reporting as safety measures may have disproportionate adverse impacts on the human rights of vulnerable groups if not addressed.

### Scenario 6 (Human Rights Risk):

## Complexity of user reporting, language, and confidentiality

Asad is a 22-year-old refugee who is currently resettled in Turkey, under the UNHCR’s “safe third country” principle. He is waiting for his asylum status for the US to be processed. Asad is introduced by a friend to a WhatsApp group that puts potential refugees in contact with human smugglers. He contacts a smuggler and agrees to meet. However, he soon finds that instead of the promised refugee visa, the smuggler is actually involved in human trafficking and organ sale. Before falling into this trap, Asad blocks this individual and seeks to report the group and the smuggler to WhatsApp with the hope of saving others, but he is scared that the smuggler might find out that he was the one who reported them and go after him. In addition, he is worried that by reporting to WhatsApp, UNHCR might get involved and he may lose his registered refugee status.



The above scenario highlights the complexity of user reporting, which depends significantly on users’ trust in Meta’s reporting mechanisms, as well as on users’ levels of literacy, digital literacy, culture, language, age, and the context of an issue. In this scenario, Asad could not find enough information about reporting in his native language, so he did not fully understand how it worked. He was also afraid to report the smuggler on WhatsApp because he was not sure what the implication of “reporting” would be, who would receive the report, and what organizations would get involved in the process of handling the situation. In many countries where there is a lack of Meta information in local languages, reporting processes are widely misunderstood. This prevents users from reporting problematic content in many cases because they are worried reporting would involve the government or result in negative consequences for them.

**Scenario 7 (Human Rights Risk):****Metadata collection, behavioral signals, and false flags**

One technique used to identify terrorist activities in an end-to-end encrypted environment is the deployment of machine learning systems that rely on the collection of metadata and behavioral signals, such as financial activities, to identify potential terrorist activity. However, identifying terrorist activity via behavioral signals is complicated for reasons mentioned previously in this report.

Amir is a graduate student in the UK who uses Meta's messaging platforms to stay in touch with his family in Iran. His family uses VPNs to evade the Iranian government's blocking of Meta's family of apps. Amir is also a member of different cryptocurrency-related messaging groups to see how he can find a way to evade financial sanctions and receive money from his family to help pay for his living expenses. Despite being completely innocent, Amir's account is automatically flagged by Meta's classifiers because his behavior closely resembles that of a terrorist group member.



The above example illustrates the complexity of user behavior and habits in different contexts. During BSR's conversation with various experts, they raised concerns about relying on machine learning-enabled techniques, believing that an overreliance on "behavioral signals" to train and optimize ML-enabled systems might lead to a high number of false positives, disproportionately impacting groups whose local languages and user behavior are overrepresented or underrepresented. This could adversely impact some users' right to freedom of expression, access to information, dignity, equality, and nondiscrimination.

# Counterbalancing Competing Rights in End-to-End Encryption

All human rights are indivisible, interdependent, and interrelated. The improvement of one right can facilitate advancement of others; the deprivation of one right can adversely affect others.

For example, privacy is a necessary condition for the realization, promotion, and protection of many other human rights, such as rights to freedom of expression, freedom of assembly and association, freedom of movement, freedom of belief and religion, and access to remedy.

However, human rights can also come into tension with one another for legitimate reasons, as the use of end-to-end encryption in private messaging services illustrates perfectly: On the one hand, end-to-end encryption protects privacy, enhances security, and enables freedom of opinion, expression, movement, association, religion, and belief; on the other hand, end-to-end encryption can hinder some efforts to protect child rights, liberty, safety, and personal security.



We believe that rights-based methods can be deployed to define a path forward when two competing rights cannot both be achieved in their entirety. Rather than “offsetting” one right against another, it is important to pursue the fullest possible expression of both rights and identify how potential harms can be addressed.

In this assessment we have used a methodology known as “counterbalancing” to identify ways to secure the fullest possible expression of rights without unduly limiting others by applying established international human rights principles such as legitimacy, necessity, proportionality, and nondiscrimination. We make recommendations for how potential adverse human rights impacts arising from counterbalancing can be mitigated by Meta or by other actors, such as law enforcement agencies.

*Counterbalancing is a methodology that can be deployed when a company needs to navigate an approach between two competing rights, and involves enabling the fullest possible expression of human rights without unduly limiting others.*

End-to-end encryption involves many instances of competing human rights, and various “sides” of the encryption debate often argue that their case is justified because it “does more good for more people than it harms,” and vice versa.

However, this view is inconsistent with international human rights principles for three important reasons: First, the UNGPs are clear that the corporate responsibility to respect human rights includes all rights; second, there is no hierarchy of rights—no one right can be considered more important than another; and third, positive human rights impacts should not be used to “offset” adverse human rights impacts.

Since Meta cannot simply decide to privilege one right over another or choose to ignore the human rights harm facilitated by end-to-end encrypted messaging in the interest of the “greater good,” it has a responsibility to assess how to make decisions about the human rights trade-offs involved in deploying end-to-end encryption in a manner consistent with international human rights principles. One way to approach this is through counterbalancing rights.

Counterbalancing is a methodology that can be deployed when a company needs to navigate an approach between two competing rights, and involves enabling the fullest possible expression of human rights without unduly limiting others.

As was stated earlier, counterbalancing is not a part of the UNGPs, which do not focus on how companies should address instances of competing human rights. This exercise is therefore meant to be illustrative of how the international human rights framework can be used to resolve some of the most challenging debates related to end-to-end encryption. BSR’s approach to counterbalancing rights in this HRIA is shaped by the following international human rights principles:

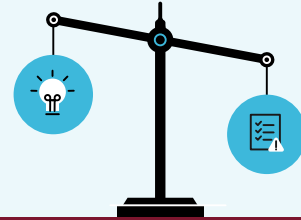
- **Reverting to principle**—Focusing on the underlying principle of the right being restricted and identifying ways to uphold the core principle, even if not the exact right.
- **Legitimacy**—Restrictions to a right must pursue an objectively legitimate purpose and address a precise threat.
- **Necessity and proportionality**—Only restricting a right when the same goal cannot be achieved by other means, and using restrictions that are the least intrusive to achieve the legitimate purpose.
- **Nondiscrimination**—Restrictions to a right must be implemented in a nondiscriminatory manner.





Counterbalancing has the potential to bring much needed clarity and nuance to the broader encryption debate, and can point toward compromises that can be reached in a manner consistent with human rights principles. Here we carry out a counterbalancing exercise with two of the most challenging cases of competing human rights in the context of end-to-end encryption—privacy against security, and privacy against the rights of children to be protected from sexual abuse and exploitation. The results of this exercise have informed our recommendations for Meta.

When considering various policy and product decisions related to end-to-end encrypted messaging that involve competing human rights, Meta can carry out a similar counterbalancing exercise to help arrive at a rights-respecting solution. However, this counterbalancing exercise can only be done with two competing rights at a

## Counterbalancing Competing Rights

When counterbalancing competing rights, it's important to utilize the following international human rights principles:



International Human Rights Principles	Questions to Ask
 <b>Reverting to principle</b>	Can the core principle of the restricted right still be upheld in different ways?
 <b>Legitimacy</b>	Is there a legitimate aim in pursuing the restriction of this right?
 <b>Necessity and proportionality</b>	Is the restriction of the right necessary or can the legitimate goal be achieved through other means? If it is necessary, is it the least intrusive way to restrict this right?
 <b>Nondiscrimination</b>	Can the restriction of the right be done in a nondiscriminatory manner?

time, when there are often many overlapping rights tensions. Because human rights are interrelated and interdependent, the rights tensions present in end-to-end encryption should not be solely considered in isolation, but collectively. This section should be read with that in mind.

It should be noted that for the purposes of this HRIA we focus only on end-to-end encryption of messaging, as distinct from other kinds of encryption of data in transit and as distinct from device or hard drive encryption (i.e., encryption

of data at rest). While many of the human rights tensions are similar for all types of encrypted data, they manifest in different ways, and so not all of our analysis here necessarily applies.

It should also be noted that this is a human rights analysis of the trade-offs between proposed solutions. There is a parallel discussion to be had about the technical integrity and soundness of these solutions, and while we reference these debates, we do not make any technical assertions of our own.



## 10.1 Counterbalancing the Right to Privacy and the Right to Security

One of the most intractable human rights tensions in the encryption debate is between the right to privacy and the right to security.

On the one hand, end-to-end encrypted messaging directly protects the privacy of users; on the other hand, some users use end-to-end encrypted messaging platforms to carry out activities that violate the security of others, such as terrorist attacks and other forms of criminal activity. This debate is often described as “privacy vs. security,” but it is also “security vs. security” because the privacy protections of end-to-end encrypted messaging also protect the bodily security of vulnerable users and rightsholders.

To see how we might counterbalance these competing rights, we look at how each of the rights in question could be justifiably limited using the counterbalancing principles.

### Option One: Limiting the privacy of all users to protect the security of others

This option covers scenarios where Meta doesn’t introduce end-to-end encryption (i.e., the status quo at the time of writing), or where end-to-end encryption is introduced with backdoors and “exceptional access.”

- **Reverting to principle:** The underlying principle of the right to privacy in the context of end-to-end encrypted messaging is that people should be able to have private conversations that allow them to express themselves freely without fear of arbitrary interference or retribution. In the interest of protecting security, many argue that companies should not implement end-to-end encryption, or that companies must weaken end-to-end encryption by creating “backdoors,” or otherwise

providing law enforcement with “exceptional access” to user communications in legitimate criminal investigations via key escrow systems, mandatory key disclosure, targeted decryption orders, or enabling law enforcement to secretly join a conversation (i.e., the “ghost proposal”).<sup>1</sup> Each of these measures constitute some form of limitation on the right to privacy. This raises the question, “can these limitations be pursued while still upholding the underlying privacy principles of end-to-end encrypted messaging?”

Because end-to-end encryption is so fundamental to the right to privacy, this principle cannot be preserved if Meta were to decide not to pursue end-to-end encrypted messaging. It also cannot be preserved in the case of “backdoors” or various forms of “exceptional access” because they could enable users’ communications to be accessed in unlawful and arbitrary ways. Exceptional access may not be arbitrary in democratic, rule-of-law-based countries where a legitimate court order is required to gain access to communications, but it would certainly be arbitrary in the growing majority of countries that lack rule of law and respect for human rights. Given Meta’s end-to-end encrypted messaging platforms are globally available, and the technical difficulties in verifying the actual location of users, Meta (and other messaging companies) cannot easily allow exceptional access “just for the good guys.” Authoritarian governments would undoubtedly seek to require their own exceptional access.

Additionally, as security experts have repeatedly explained,<sup>2</sup> creating a backdoor is creating a vulnerability in the system. Once that backdoor exists, hackers and bad actors around the world will try to access it for their own ends. This

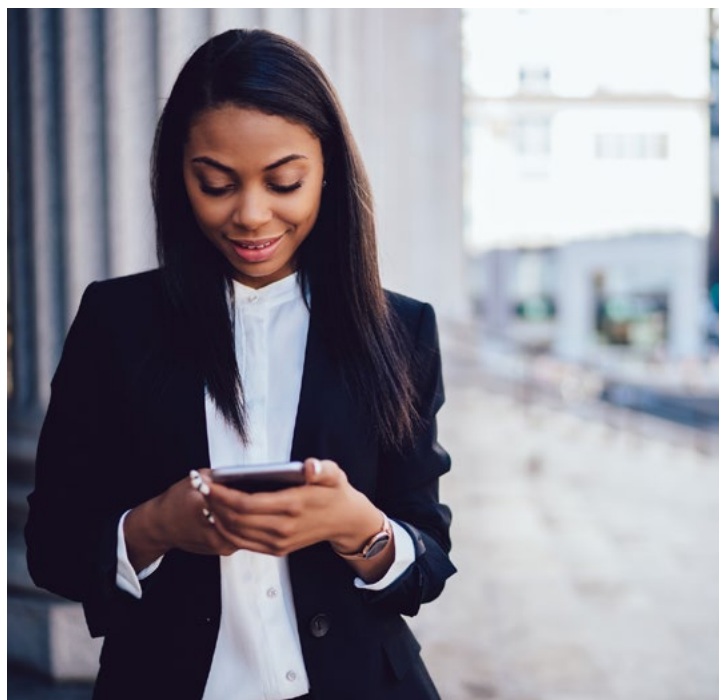
<sup>1</sup> <https://www.lawfareblog.com/principles-more-informed-exceptional-access-debate>.

<sup>2</sup> For example, <http://dspace.mit.edu/handle/1721.1/97690>, <https://www.lawfareblog.com/open-letter-gchq-threats-posed-ghost-proposal>, <https://www.justsecurity.org/53316/criminalize-security-criminals-secure/>, <https://www.washingtonpost.com/news/powerpost/paloma/the-cybersecurity-202/2018/06/11/the-cybersecurity-202-we-surveyed-100-experts-a-majority-rejected-the-fbi-s-push-for-encryption-back-doors/5b1d39eb1b326b6391af094a/>, <https://www.thirdway.org/report/weakened-encryption-the-threat-to-americas-national-security>, <https://www.newamerica.org/weekly/encryption-backdoors-put-more-risk-you-might-think/>.

was demonstrated recently by the SolarWinds cyberattack in which hackers were able to steal US Treasury Department encryption keys.<sup>3</sup> This vulnerability means that backdoors can result in additional violations of privacy and security of users, as they could ultimately be exploited by the same criminal and terrorist organizations it seeks to target.

- **Legitimacy:** Limiting privacy in order to protect the security of others by detecting and preventing crime is a legitimate aim. In fact, this is why the right to privacy as laid out in the ICCPR is inherently limited to “unlawful and arbitrary” interference. Society has long accepted that law enforcement can justifiably access private communications as part of criminal investigations as long as there is appropriate legal authorization and sufficient safeguards. However, governments with questionable human rights records regularly request access to private communications in ways that are not consistent with the rule of law or use vague and overbroad laws under the guise of protecting national security when their real goal is to surveil human rights defenders or political dissidents.

- **Necessity and proportionality:** Much has been written in the human rights field about the necessity and proportionality of proposed limits to encryption,<sup>4</sup> most notably by former UN Special Rapporteur for Freedom of Expression, David Kaye. Some policymakers have called for encryption bans, which he argues are not necessary or proportionate to achieving public security “because they deprive all online users in a particular jurisdiction of the right to carve out private space for opinion and expression, without any particular claim of the use of encryption for unlawful ends,” and because “measures that impose generally applicable restrictions on massive numbers of persons, without a case-by-case assessment, would almost certainly fail to satisfy proportionality.”<sup>5</sup> As a private



company, a decision by Meta to not implement end-to-end encryption would not be the same as a government ban. However, as discussed previously, it may lead to Meta “contributing to” human rights harm for failing to protect the privacy and security of users.

If Meta were to acquiesce to government calls to weaken its end-to-end encrypted messaging systems via backdoors it would also not be necessary or proportionate to the aim of fighting crime because as Kaye argues, governments “have not demonstrated that criminal or terrorist use of encryption serves as an insuperable barrier to law enforcement objectives,” and because the privacy and security impacts of a backdoor that falls into the wrong hands would be widespread and indiscriminate, and disproportionately affect all users and rightsholders.<sup>6</sup> Kaye also generally argues that key escrow systems are not necessary or proportionate because “the vulnerabilities inherent in key escrows render them a serious threat to the security to exercise the freedom of expression,” and that mandatory key disclosure are also not necessary and proportionate because

<sup>3</sup> <https://www.axios.com/solarwinds-hack-treasury-email-accounts-breached-e6a24240-2795-4c09-9056-b53f20e47f37.html>.

<sup>4</sup> For example: <https://www.amnestyusa.org/reports/encryption-a-matter-of-human-rights/>.

<sup>5</sup> <https://www.undocs.org/A/HRC/29/32>.

<sup>6</sup> *Ibid.*, para. 42.

they would enable access to an entire set of messages encrypted by a specific key, rather than just those pertinent to the investigation.”<sup>7</sup>

Targeted decryption via judicial order and the ghost proposal (which is similar in scope to targeted decryption in that it is focused on specific targets and communications, see citation above) could be considered necessary and proportionate if Meta limited disclosure to specific communications as part of a legitimate law enforcement investigation. According to Kaye, targeted decryption orders “should be based on publicly accessible law, clearly limited in scope focused on a specific target, implemented under independent and impartial judicial authority, in particular to preserve the due process rights of targets, and only adopted when necessary and when less intrusive means of investigation are not available. Such measures may only be justified if used in targeting a specific user or users, subject to judicial oversight.”<sup>8</sup>

However, although targeted decryption could be considered necessary and proportionate from a human rights perspective, it would require companies to weaken or break encryption, essentially becoming a “backdoor” that suffers from the same privacy and security risks mentioned above. Related, the ghost proposal would undermine the authentication systems used in end-to-end encrypted messaging to verify that the other users in a conversation are who they say they are and that the conversation is secure, therefore weakening protections and undermining user trust.<sup>9</sup>

- **Nondiscrimination:** A decision not to pursue end-to-end encryption and encryption backdoors would affect all users. However, targeted decryption, key disclosure orders, and use of the ghost proposal could be pursued by law enforcement in a discriminatory manner. In

addition, vulnerable communities who do not have the resources, know-how, or technical standards on their phones to migrate to more secure services would likely be disproportionately impacted by this option.

### Option Two: Inhibiting the protection of security and bodily integrity of some users in order to protect the privacy of all users<sup>10</sup>

This option covers the scenario where Meta does introduce end-to-end encryption, without backdoors or exceptional access.

- **Reverting to principle:** The underlying principle of security and bodily integrity in this context is keeping people safe from terrorist attacks and crime. So how can Meta help government and law enforcement officials do that in an end-to-end encrypted context? Law enforcement officials often argue they will not be able to prosecute cases without access to message content, but criminal investigations always require multiple pieces of evidence, of which message content is just one. In the absence of message content, Meta can provide law enforcement with other kinds of non-encrypted information that is useful for preventing and investigating crimes, such as metadata, user reports, public platform content, or behavioral indicators arising from the linking of private messaging with public facing accounts. Law enforcement around the world often lack the tools and skills required to identify and make use of the massive amount of digital evidence at their disposal, so it will be important for Meta to help law enforcement understand what kind of information they can provide and how to make sense of it (e.g., via training or innovative investigative tools) for the underlying principles of security to be maintained.

<sup>7</sup> Ibid, para. 43.

<sup>8</sup> Ibid, para. 45.

<sup>9</sup> See <https://www.lawfareblog.com/open-letter-gchq-threats-posed-ghost-proposal> for more information on the security and trust-related risks of the ghost proposal.

<sup>10</sup> Note that, as explained above, the rights conflicts related to end-to-end encryption are not binary. In this case there is not just inhibiting of privacy, there is also inhibiting of bodily security rights due to the privacy protections end-to-end encryption provides. However, the counterbalancing exercise is meant to explore conflicts between two rights at a time for reasons of complexity. We therefore explore the right to privacy in this case, as it is the enabling right.

- **Legitimacy:** The protection of privacy that is directly enabled by end-to-end encryption is a legitimate aim, and addresses numerous precise threats to privacy that were explored earlier in this report. Importantly, the existence of end-to-end encryption does not in and of itself limit the right to security and bodily integrity; rather, it inhibits some efforts to protect those rights.
- **Necessity and proportionality:** The only way to truly protect the privacy of communications in the face of the various threats mentioned previously in this report (such as surveillance or bodily harm) is via end-to-end encryption. It is worth reiterating here that removing the possibility of law enforcement gaining access to message content does not fully prevent law enforcement's ability to investigate a crime, as messaging content is just one piece of the large investigative puzzle. Regardless of Meta's expansion of end-to-end encryption, law enforcement agencies

*In the case of the classic privacy against security debate, counterbalancing suggests that Meta should favor the privacy rights of all users. End-to-end encryption directly enables privacy, therefore the principles of privacy cannot be meaningfully upheld by banning end-to-end encryption, weakening it with backdoors or any form of "exceptional access."*

today benefit from vastly more data and advanced data analysis capabilities than in the past, and have more data available for analysis than ever before. However, because end-to-end encryption does ultimately have some negative impact on law enforcement's ability to protect security by carrying out investigations using traditional techniques, companies should mitigate this impact by proactively working with law enforcement to access, understand, and utilize other kinds of data in ways that are lawful, necessary, proportionate, and nondiscriminatory.

- **Nondiscrimination:** Because the negative impacts of end-to-end encryption on this area of security do not disproportionately impact certain groups, nondiscrimination is not relevant in this case.

## Conclusion

In the case of the classic privacy against security debate, counterbalancing suggests that Meta should favor the privacy rights of all users. End-to-end encryption directly enables privacy, therefore the principles of privacy cannot be meaningfully upheld by banning end-to-end encryption, weakening it with backdoors or any form of "exceptional access."

Additionally, restrictions on encryption are neither necessary nor proportionate to the goal of public safety and are likely to be implemented by nondemocratic countries in ways that are arbitrary, and therefore illegitimate, as well as discriminatory. Because security rights would be somewhat limited in this case, companies should seek to mitigate the impact on security and uphold the principle of keeping people safe from terrorist attacks and crime by working proactively with law enforcement agencies to provide them with data that is useful in investigations (subject to proper legal process), and the tools to make sense of it.

## 10.2 Counterbalancing the Right to Privacy of All Users with the Rights of Children to Be Protected from Sexual Abuse and Exploitation

Another of the most difficult human rights tensions in the encryption debate is between the right to privacy of all users and the protection of children from sexual abuse and exploitation. The privacy protections of encryption protect the privacy rights of everyone (including children), but it is also known that some users utilize end-to-end encrypted messaging platforms to sexually exploit children by grooming them for trafficking or other forms of exploitation, as well as for the production, dissemination, and viewing of child sexual abuse material.

Child protection advocates from both the technical and nontechnical communities have been researching methods to protect children in a messaging context, such as innovations in user reporting and the use of behavioral signals to identify and interrupt inappropriate interactions between adults and children.

One of the most heated debates has been about the detection of CSAM in end-to-end encrypted contexts. Researchers have been developing various client-side scanning methods with the goal of identifying an approach that can preserve user privacy as much as possible while enabling companies to detect and report the sharing of CSAM.

Privacy advocates have largely been opposed to the use of client-side scanning in any form, arguing that these methods amount to an encryption “backdoor” and censorship tool.<sup>11</sup> Meta’s position to date has been similar, in line with the broader definition of end-to-end encryption meaning all information about the content of a message is known only to the sender and intended recipients (see Section 5.3). This has created a conflict with competing equities for which there has been little room for compromise

because both the child protection advocates and privacy advocates see their respective issues as paramount.

To explore how these competing rights might be counterbalanced, we examine how each of the rights in question could be justifiably limited using counterbalancing principles. Note that this section does not make any assertions about the technical merits of various CSAM detection proposals, but rather discusses the related human rights trade-offs. Discussion of potential technical uncertainties and risks raised by technical experts can be found in the following section.

### Option One: Limiting the privacy of all users to protect children from sexual abuse and exploitation

This option covers scenarios where Meta does not expand end-to-end encryption (i.e., the status quo at the time of writing), or where end-to-end encryption is introduced with some form of scanning for CSAM.

- **Reverting to principle:** The underlying principle of the right to privacy in the context of end-to-end encryption is that people should be able to have private conversations that allow them to express themselves freely without fear of arbitrary interference or retribution. Because encryption is so fundamental to the right to privacy today (see Section 7.1), this principle cannot be preserved without providing end-to-end encryption, by banning end-to-end encryption, or by weakening or compromising the cryptographic integrity of end-to-end encryption.

However, Meta could in theory preserve the principles underlying privacy while pursuing nascent client-side scanning approaches to

<sup>11</sup> For example, see <https://cdt.org/wp-content/uploads/2021/08/CDT-Outside-Looking-In-Approaches-to-Content-Moderation-in-End-to-End-Encrypted-Systems.pdf>.

enable the detection of CSAM in end-to-end encrypted messaging without cryptographically weakening or breaking encryption. Thus far, the only approach proposed that may not undermine the cryptographic integrity of end-to-end encryption is homomorphic encryption, because it would allow data to be “processed” while it is still encrypted. This means it could uphold the underlying principle of ensuring conversations between people remain private and free of arbitrary interference.

However, homomorphic encryption is not compatible with the broad definition of end-to-end encryption—i.e., the idea that there should be absolutely nothing in the middle of a conversation between the two end points. It is also not technically feasible using today’s computing power (see Section 11), and this means there is currently no feasible approach to detecting CSAM in an end-to-end encrypted environment that would not undermine the principle of privacy.

- **Legitimacy:** Limiting privacy in order to protect children from sexual abuse and exploitation is a legitimate aim, particularly since many of the instances of child sexual abuse and exploitation that can occur via end-to-end encrypted messaging are nearly universally recognized to be crimes. Thus, detecting content like CSAM is legitimate and may be justified and does not constitute an arbitrary or unlawful interference with privacy.
- **Necessity and proportionality:** Restricting the privacy of all users by banning, weakening, or choosing not to adopt end-to-end encryption in order to protect children from sexual abuse and exploitation is not necessary. Meta can deploy alternative measures to prevent this harm from occurring, such as analyzing user behavior with metadata that reveal indicators of child exploitation, utilizing classifiers to identify users seeking to exploit children, and facilitating increased user reporting. However, it is important to remember that child sexual

abuse and exploitation online takes many forms, with associated harms, and the only way to fully address the harm caused by the distribution of CSAM is through detection and blocking. Although Meta is using behavioral signals to detect and prevent CSAM distribution groups from forming, these measures cannot fully prevent the sharing of CSAM. Measures designed to prevent the sexual exploitation of children from occurring in the first place also do not address the challenge of existing CSAM. If it becomes technically feasible to detect CSAM using client-side scanning methods that maintain the cryptographic integrity of end-to-end encryption and do not create other adverse human rights impacts, then banning, choosing not to implement, or weakening end-to-end encryption would not be necessary.

Banning, weakening, or choosing not to adopt end-to-end encryption in order to protect children from exploitation would also not be proportionate because it would pose an undue burden on the privacy rights of all users. Indeed, the alternative measures discussed above may not only constitute less privacy-restrictive ways to prevent grooming and inappropriate interactions with children in an end-to-end encrypted environment, but they may also hold the potential to be more effective and preventative over time.

- **Nondiscrimination:** Banning or not implementing end-to-end encryption, or implementing techniques such as homomorphic encryption, would affect all users. However, unintentionally discriminatory outcomes could occur due to the possibility that even if they become technically feasible, some of the CSAM detection techniques may still be too computationally intensive for low-end devices or in places with low bandwidth. In addition, vulnerable communities who do not have the resources, know-how, or technical standards on their phones to migrate to more secure services would likely be disproportionately impacted by any decisions not to implement or to weaken end-to-end encryption.

## Option Two: Limiting the right of children to be protected from sexual abuse and exploitation to protect the privacy of all users

This option covers the scenario where Meta does introduce end-to-end encryption, without any form of scanning for CSAM.

- **Reverting to principle:** In the context of private messaging services, the underlying principle of the right of children to be protected from sexual abuse and exploitation means ensuring children are protected from grooming and inappropriate interactions with adults while using private messaging systems. It also means protecting victims from the privacy violations and revictimization they suffer when CSAM is shared. However, it is important to recognize that children's right to privacy online is an important component of protecting them from exploitation. A recent UNICEF report states that a child's right to privacy and protecting children from abuse and exploitation must be equally upheld; that privacy cannot be viewed as secondary.<sup>12</sup>

In order to uphold these principles in an end-to-end encrypted context, Meta can endeavor to create safe environments for children on their messaging platforms through non-content based forms of preventing and detecting child exploitation—such as metadata analysis and the use of behavioral classifiers (e.g., identifying cross-platform signals made by users seeking to exploit children, such as contacting minors in a different city, and analyzing user behavior for indicators of child exploitation), and improved user reporting. They can also proactively work with child protection organizations and law enforcement on child exploitation cases to help them handle the volume of data they receive from CSAM reports and turn it into actionable information. Some public and / or unencrypted content data, such as user profiles and pictures, can also be used.

One resource companies can turn to for guidance is UNICEF's Guidelines for Industry on Child Online Protection, which outline five key areas for companies to protect and promote children's rights online: (1) integrating child rights considerations into all appropriate corporate policies and management processes; (2) developing standard processes to handle child sexual abuse material; (3) creating a safer and age-appropriate online environment; (4) educating children, parents, and teachers about children's safety and their responsible use of technology; and (5) promoting digital technology as a mode for increasing civic engagement.<sup>13</sup>

While such measures are important for keeping children safe on messaging services and preventing and addressing many forms of child sexual abuse and exploitation, the only way known to date to fully prevent the sharing of CSAM and address the associated human rights harm associated is to detect, block, and report it. There is no way to uphold the principle of this specific component of the right of children to be protected from sexual abuse and exploitation in an end-to-end encrypted environment without implementing some form of client-side scanning.

- **Legitimacy:** The protection of privacy for all users that is directly enabled by end-to-end encryption is a legitimate aim and addresses numerous precise threats to privacy, as has been explored throughout this HRIA. Importantly, the existence of end-to-end encryption does not in and of itself limit children's right to be protected from sexual abuse and exploitation, but it does make the detection of sexual abuse and exploitation of children via activities like grooming and CSAM distribution more difficult to detect.
- **Necessity and proportionality:** The only way to truly protect the privacy of communications in the face of the various global threats mentioned previously is through end-to-end encryption. Nevertheless, restricting the right of children to be free from sexual abuse and exploitation

<sup>12</sup> [https://www.unicef-irc.org/publications/pdf/Encryption\\_privacy\\_and\\_children%E2%80%99s\\_right\\_to\\_protection\\_from\\_harm.pdf](https://www.unicef-irc.org/publications/pdf/Encryption_privacy_and_children%E2%80%99s_right_to_protection_from_harm.pdf).

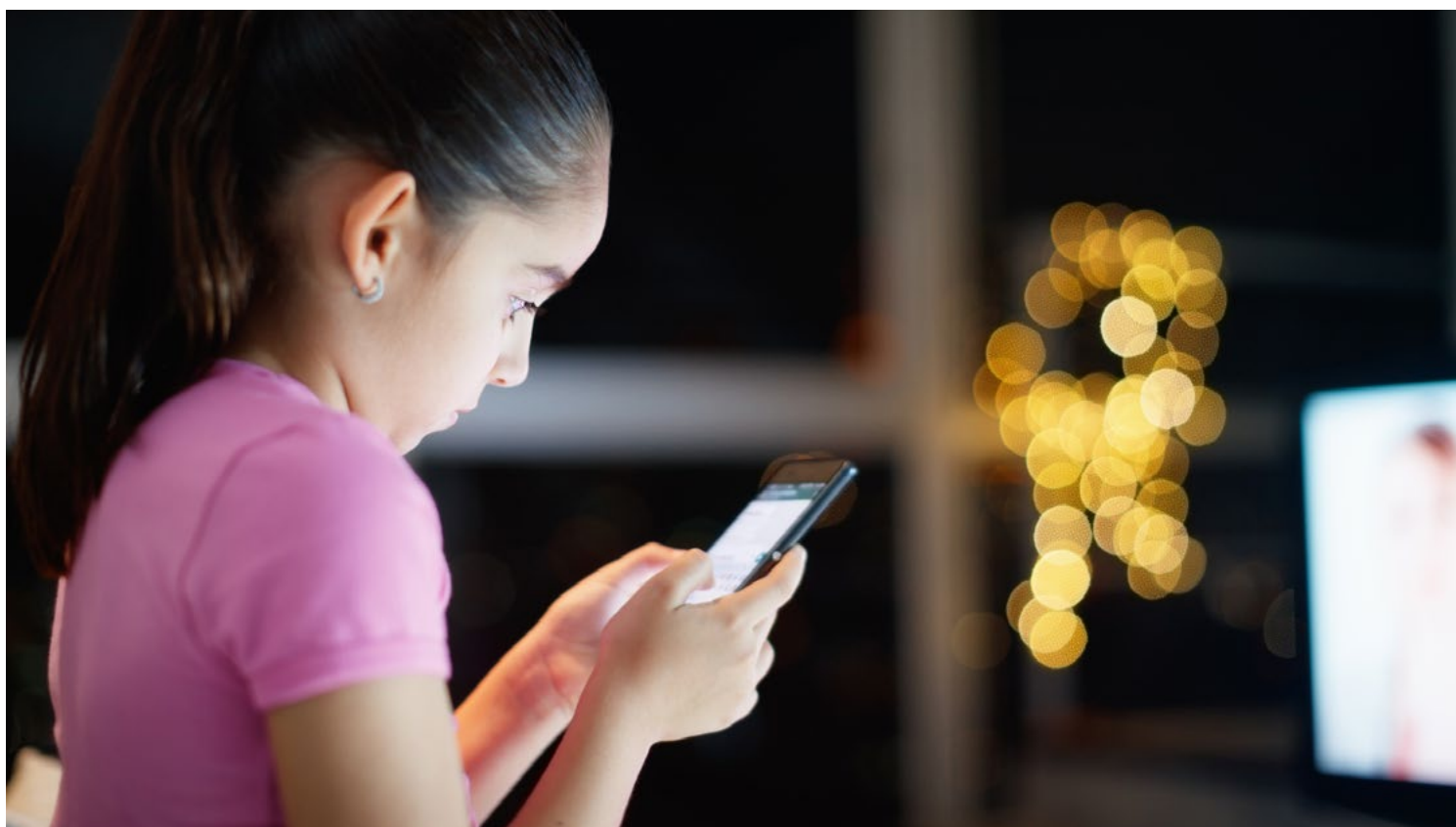
<sup>13</sup> <https://www.unicef.org/media/66616/file/Industry-Guidelines-for-Online-ChildProtection.pdf>.

in favor of protecting broader user privacy via end-to-end encryption without any form of child exploitation detection would not be necessary or proportionate if there were technically feasible methods that have less of a negative impact on these rights—for example, the possibility of detecting CSAM while maintaining the cryptographic integrity of end-to-end encryption via homomorphic encryption. However, these methods are not yet technically feasible at scale using available computing power today, and any method that technically weakens or breaks encryption would constitute a disproportionate restriction on the right to privacy.

- **Nondiscrimination:** Because the negative impacts of end-to-end encryption on the right of children to be free from sexual abuse and exploitation would not disproportionately impact certain groups over others, nondiscrimination is not relevant in this case.

## Conclusion

In the debate between the privacy rights of all users and the right of children to be free from sexual exploitation and abuse, counterbalancing suggests two possible conclusions. The first potential conclusion is a compromise that results in a slight limitation of each right by deploying approaches that can maintain the cryptographic integrity and most of the privacy protections of end-to-end encryption through methods of CSAM detection (such as partial or fully homomorphic encryption). In this scenario, counterbalancing would suggest that if such measures are technically feasible, and do not create other adverse human rights impacts, they should be pursued because they uphold the principles of privacy while representing the least restrictive burden on children’s right to be free from sexual abuse and exploitation.





However, since these methods of CSAM detection are currently not technically feasible, and may pose other adverse human rights impacts, the next best option suggested by counterbalancing would be for Meta to continue investing in efforts to create safe online environments for children on its platforms, such as seeking to improve the use of behavioral signals and user reports to detect the formation of CSAM distribution groups, grooming, and other forms of exploitation, preventing abusers from rejoining Meta platforms, methods to guide and influence user behavior, and proactive work with child protection organizations and law enforcement to help them process and make sense of the volume and variety of data they receive as part of child exploitation cases.

Beyond this, an option proposed by some child rights organizations would be for Meta to limit end-to-end encrypted messaging to users 18 and older. This solution might help make it easier to prevent and detect inappropriate interactions with children, such as grooming and solicitation. However, it is unlikely to address the issue of CSAM shared among adults. It is unclear if this solution would be technically feasible in the context of messaging systems that can cross-communicate, and it would likely also be difficult to verify users' ages if they cannot be connected to public platform accounts. This option would also deny the privacy protections of end-to-end encrypted messaging to children, and especially teenagers, in contexts where it might be vital to their physical safety. Given the promise of prevention methods in an end-to-end encrypted environment this option may also not be necessary.

Another option, which has been proposed by some child protection organizations, would be for Meta to delay its expansion of end-to-end encryption until client-side scanning techniques that maintain cryptographic integrity are technically feasible. However, this decision would also come with human rights risks associated with the lack of privacy protections.

*Although implementing encryption-respecting client-side scanning—which is not currently technically feasible at scale—may not impose an undue burden on the privacy rights of users, it does come with other significant human rights risks that may be impossible to address.*

Importantly, the tension between the right of children to be free from sexual exploitation and abuse and the right to privacy of all users does not exist in a vacuum. Although implementing encryption-respecting client-side scanning—which is not currently technically feasible at scale—may not impose an undue burden on the privacy rights of users, it does come with other significant human rights risks that may be impossible to address. These risks are explored in the following section.

# The Human Rights Trade-offs of ‘Client-Side Scanning’ for Content Moderation in an End-to-End Encrypted Environment

Methods such as client-side scanning of a hash corpus, trained neural networks, and multiparty computation including partial or fully homomorphic encryption (often collectively referred to as “perceptual hashing” or “client-side scanning,” although some can also be server-side) have all been suggested as solutions to enable Meta to continue to identify, remove, and report CSAM in an end-to-end encrypted environment.

In this assessment we use the term “client-side scanning” as a catch-all for any approach to detecting content in messaging.



These methods are nascent, having largely been tested in academic settings, and are therefore unproven in real-world context. However, this changed recently with Apple’s August 2021 announcement that it would be rolling out client-side scanning in the US to detect CSAM in photos on users’ devices prior to upload to iCloud, and to detect and offer prompts when sexually explicit photos are sent or received on children’s iMessage accounts.<sup>1</sup> The announcement received a range of feedback, with child protection groups praising Apple for the move, while technologists and digital

<sup>1</sup> <https://techcrunch.com/2021/08/05/apple-icloud-photos-scanning/>.

*This debate is a microcosm of a much larger debate about content moderation in both private messaging services in general and end-to-end encrypted messaging in particular.*

rights advocates expressed concern about the lack of technical consensus, risks to technical integrity, the potential for adversarial manipulation of the tool, and the risk that it opened the door to government censorship and surveillance. Apple ultimately slowed the roll-out of these features.

The conflicting nature of the public response to the Apple announcement was no surprise given the tensions inherent in the history of the encryption debate. However, as we noted in the call out box earlier in this report, the arguments in the Apple case cannot be neatly copy-pasted onto the debate about whether companies should detect CSAM in an end-to-end encrypted messaging environment.

This debate is a microcosm of a much larger debate about content moderation in both private messaging services in general and end-to-end encrypted messaging in particular. Although we conclude in our counterbalancing analysis that client-side scanning methods that preserve the cryptographic integrity of end-to-end encryption (such as homomorphic encryption) could in theory be a justifiable limitation of privacy, there are potential knock-on effects of a decision to pursue CSAM detection on other human rights that need to be carefully considered.

In theory, hash-based approaches (such as the one proposed by Apple) to detect CSAM could also be used to detect, block, and/or remove many other kinds of objectionable content, such as nonconsensual intimate images, hate speech, and terrorist content. This means that the debate about client-side scanning for CSAM detection raises several other challenging dilemmas for which there are no easy answers. In this section we review the following:

1. There is no consensus on where to draw the line on content moderation in a private messaging context.
2. The technical feasibility, integrity, and resiliency of client-side scanning methods is uncertain.
3. There is a “slippery slope” risk. Choosing to moderate content like CSAM could lead to regulatory requirements that Meta moderate broader categories of content.

*Although we conclude in our counterbalancing analysis that client-side scanning methods that preserve the cryptographic integrity of end-to-end encryption (such as homomorphic encryption) could in theory be a justifiable limitation of privacy, there are potential knock-on effects of a decision to pursue CSAM detection on other human rights that need to be carefully considered.*

**Dilemma 1: There is no consensus on where to draw the line on content moderation in a private messaging context.**

Traditionally, messaging services—both SMS and internet-based—have been considered private territory. When people send messages to each other, they generally expect that those messages will remain between the intended parties. This is quite different from a public social media platform, where users post content knowing it is not private.

Meta has historically treated its messaging platforms as private domains. With the exception of particularly egregious content such as CSAM, Meta does not actively enforce its various content policies on its messaging platforms; instead, it relies on users to block and report interactions they object to. It is reasonable for content and acceptable use policies on private messaging services to be quite different than those on more public social media platforms, owing to the very different nature of the services being provided.

Increasingly, however, messaging platforms are no longer just private domains.<sup>2</sup> The existence of large WhatsApp groups means that private messaging can sometimes feel more like a quasi-public space to some users, with many of the same challenging content issues seen on traditional social media platforms—albeit without the same level of discoverability and searchability, without algorithmic promotion of content, and for a very small percentage of overall chats on WhatsApp. The impacts of viral misinformation, hate speech, and other types of problematic content on WhatsApp has led to some calls to impose the same kinds of content moderation standards in the messaging space that are implemented on social media platforms.

But messaging platforms are still largely private spaces, and many legitimately argue that moderation of anything other than the most egregious types of content would be an unnecessary and disproportionate infringement on privacy and free expression.

Hash-based systems that could operate in an end-to-end encrypted environment, such as those often used to detect CSAM and terrorist content, are a rather blunt content moderation tool. They rely on having an exact or near exact copy of the content that has been hashed—whether it be an image, video, or text—in order to identify that same piece of content in future messages, and this makes dealing with nuanced content very difficult. By contrast, public platforms are able to use human reporting, human review, and automatic detection via ML classifiers to identify problematic content, and as a result make enforcement decisions about content with a greater understanding of context and nuance. Hate speech, which often uses coded language and is highly contextual, is one example where this kind of analysis is necessary.

While hash-based systems are good at identifying clear-cut content—i.e., a specific image, video, or link—they cannot be used to identify nuanced types of problematic content that the system does not already have on record as part of the database. As a result, seeking to moderate broad and nuanced types of problematic content in end-to-end encrypted messaging, such as hate speech or harmful dis/misinformation, would likely result in removing far too much legitimate content and/or having a high error rate.

<sup>2</sup> There are several similarities between private messaging platforms and mass emails or mass SMS texts. This review does not cover the use of emails or SMS, but the question of content moderation in a messaging context is one for the broader industry. Facebook has an opportunity to play an active role in this conversation and emphasize the importance of addressing it from a human rights perspective.

*From a human rights perspective, only content that always and clearly constitutes a severe human rights violation when shared should be proactively moderated in an end-to-end encrypted messaging context.*

This would be compounded by the fact that remediation in a messaging context is inherently flawed. If content is blocked from being sent, not only may it be difficult for Meta to verify the accuracy of the block, but the remedies for a wrongful enforcement action, such as allowing the content to be shared or restoring an account that had been unjustly suspended, could come too late to matter to the affected user.

The limitations of hash-based content moderation and the risks of over-enforcement and errors leads us to conclude that, from a human rights perspective, only content that always and clearly constitutes a severe human rights violation when shared should be proactively moderated in an end-to-end encrypted messaging context. We believe that this content is limited to CSAM and nonconsensual intimate images because both constitute live violations of people's privacy and revictimization when shared, and both are clear-cut types of content<sup>3</sup> that can be easily included in a hash database. While many other types of content—such as hate speech or incitement to violence—may also constitute a human rights violation, they are too nuanced and contextual to be accounted for in a hash-based system.

### **Dilemma 2: The technical feasibility, integrity, and resiliency of client-side scanning methods is uncertain**

As mentioned previously, several specific client-side scanning solutions have been proposed to enable messaging services to identify, remove, and report objectionable content such as CSAM, but the only method proposed thus far that may not undermine the cryptographic integrity of end-to-end encryption is homomorphic encryption. However, homomorphic encryption is still fairly nascent and is not currently technically feasible at scale. There are also concerns about its technical integrity and resiliency in a real-world context.

*Because homomorphic encryption is still fairly nascent, there are several concerns about its technical feasibility at scale, as well as its technical integrity and resiliency in a real-world context.*

Homomorphic encryption is far too computationally intensive to implement on a large-scale messaging platform today, even for high-end mobile devices, and therefore is not technically feasible. For example, Meta's own research of a homomorphic encryption approach found that it would take around 20 million seconds (over seven months) to run on each message. Additionally, any client-side scanning solution would need to work on low-end devices, which are used by a large percentage of Meta users, for it to be effective. Eventually homomorphic encryption (and perhaps other methods that preserve cryptographic integrity) will become technically feasible as research advances and computational power increases, but it is unclear when that will be.

<sup>3</sup> However, it is important to note that nonconsensual intimate imagery often requires context and / or confirmation, and so is not as clear-cut as CSAM, which is always violating regardless of context.

Security and cryptography experts have also raised concerns about the technical integrity and resiliency of any hash-based client-side scanning systems deployed in a real-world context. For example, there is the risk that bad actors may take advantage of the technical vulnerabilities of these solutions to game the system, for example by creating false negatives to enable violating content to pass or by creating false positives to erroneously flag non-violating content, or by using unofficial clients to deactivate the code running client-side scanning. These concerns have been brought to the forefront with the Apple announcement, where security experts and cryptographers have taken issue with the closed nature of the launch, the proprietary nature of Apple's hashing algorithm, error rates, the risk of collisions (i.e., duplicate hashes produced for the same image, which means non-CSAM could be erroneously flagged as CSAM), and the ways the system could be manipulated adversarially.<sup>4</sup> Although many of these technical risks may be able to be mitigated, this underscores the problems with implementing a complex and untested system without engaging the technical community at large and without publicly proving its integrity and viability.

**Dilemma 3: There is a “slippery slope” risk. Choosing to moderate content like CSAM could lead to regulatory requirements that Meta moderate broader categories of content**

Even if cryptographic integrity-maintaining client-side scanning in end-to-end encrypted messaging were technically feasible for detecting CSAM, there is a risk that this capability could be abused by governments to require Meta to block and report legitimate content that a government dislikes. This “slippery slope” risk has emerged consistently over the history of the internet as content moderation expectations have grown, and it was a key argument of the opponents of

PhotoDNA, the hash-based system developed in 2009 that is still widely used to detect CSAM across the internet. PhotoDNA skeptics argued that the existence of the technology could lead to requirements for companies to take down all kinds of content, and that this could compromise the free and open nature of the internet.

The slippery slope risk then did not come to pass as skeptics had feared. However, the political context has changed significantly in the past two decades, and the risk is very different today than it was when PhotoDNA was invented. Government regulation of online content around the world has grown enormously, both in the legitimate pursuit of safe and rights-respecting online spaces and in the illegitimate pursuit of censorship and oppression, and as a result the risk of government mandated content moderation of end-to-end encrypted messaging has increased.

In fact, the seriousness of the slippery slope risk has been the central argument behind those opposed to Apple's new child safety features. Many have argued that the increase in requirements (and proposed requirements) from governments to track, block, and report content in countries known for suppressing dissent make it reasonable for Meta to fear that implementing client-side scanning for CSAM in the current regulatory environment would show that content moderation in an end-to-end encrypted environment is possible, and lead to governments requiring Meta to scan their own hash databases of content they dislike. Even some who had previously proposed client scanning approaches have subsequently altered their view for these reasons.<sup>5</sup>

These demands could include all kinds of content, including some unsuited to hash-based moderation, such as hate speech, and some more hash-friendly content, such as URLs, images, or videos that are critical of a government. This would undoubtedly lead to the unjust restriction

<sup>4</sup> For example, see: [https://twitter.com/matthew\\_d\\_green/status/1423071186616000513](https://twitter.com/matthew_d_green/status/1423071186616000513), <https://twitter.com/jonathanmayer/status/1427974991199543300?s=20>, and <https://www.theverge.com/2021/8/18/22630439/apple-csam-neuralhash-collision-vulnerability-flaw-cryptography>.  
<sup>5</sup> See <https://www.washingtonpost.com/opinions/2021/08/19/apple-csam-abuse-encryption-security-privacy-dangerous/>, <https://www.nytimes.com/2021/08/11/opinion/apple-iphones-privacy.html>, <https://www.eff.org/deeplinks/2021/08/apples-plan-think-different-about-encryption-opens-backdoor-your-private-life>, <https://www.accessnow.org/apple-encryption-expanded-protections-children/>.

*If the implementation of client-side scanning solely to detect CSAM—a legitimate aim—would likely result in a significant restriction of freedom of expression and other rights, then client-side scanning should not be pursued.*

of both privacy and the freedom of expression rights of users, and could erode the safe space that end-to-end encrypted messaging provides for people living in authoritarian countries, particularly for vulnerable groups. In addition, there are risks related to how hackers and companies (e.g., through insider threats) may be able to exploit such approaches by altering a hash database and putting users at risk.

It is possible that the slippery slope risk would never come to pass, or would only be pursued by governments that Meta can easily resist or ignore, and therefore a decision to protect children from sexual abuse and exploitation by detecting CSAM via client-side scanning would have minimal downstream adverse human rights impacts. However, due to the extent of the risk in the current regulatory climate, if client-side scanning were technically feasible today, it would be reasonable for Meta to decide not to implement it for fear that it would show that content moderation in an end-to-end encrypted environment is an option and result in the slippery slope risk becoming a reality.

The slippery slope risk may change over time as regulatory trends and content moderation debates evolve, and the risk should therefore be weighed by Meta when client-side scanning approaches that maintain cryptographic integrity become technically feasible. **If the implementation of**

**client-side scanning solely to detect CSAM—a legitimate aim—would likely result in a significant restriction of freedom of expression and other rights, then client-side scanning should not be pursued.** In this case, the privacy and freedom of expression violations enabled and incentivized by client-side scanning for CSAM detection would constitute a disproportionate restriction on the rights of all users.

## Conclusion

It is important to consider the complex nature of making trade-offs when assessing the possibilities of content moderation in an end-to-end encrypted environment. The use of client-side scanning could be a helpful tool to protect people from harm, but the benefits could be undermined in scenarios where client-side scanning is abused, weakens end-to-end encryption, or leads to a regulatory slippery slope in Meta's relationship with governments.

Due to both the technical complexity and the human rights trade-offs, efforts to develop and implement client-side scanning should involve multi-stakeholder participation and dialogue, and be as open and transparent as possible. Any solution should also be subject to dedicated human rights due diligence before implementation to examine the potential impacts of specific design choices and contextual factors. It is important to underscore that there are no easy answers to addressing the trade-offs, and that there are legitimate rights-based arguments both for and against client-side scanning. BSR has sought to illuminate some potential rights-based paths toward resolving those tensions, but the fast moving nature of the slippery slope risk makes that challenging.

Meta can and should proactively seek to limit the slippery slope risk by advocating for rights-respecting data protection and content moderation regulations around the world, and specific recommendations for Meta related to content moderation in its expansion of end-to-end encryption can be found in the following section.

# Recommendations

The table below contains recommendations for how Meta should avoid, prevent, and mitigate the potential adverse human rights impacts arising from the expansion of end-to-end encryption, while also maximizing the beneficial impact end-to-end encryption will have on human rights. Each recommendation is accompanied by an explanation based on the insights gained during this assessment and the expectations of the UNGPs, Global Network Initiative (GNI) commitments, and other international human rights principles.

It should be noted that these recommendations do not make any technical assertions about encryption or mitigation tactics beyond those communicated to us by Meta or external stakeholders. It also does not cover all the human rights implications of cross-app communication between Messenger, Instagram DMs, and WhatsApp, though elements of cross-app communication that are directly relevant for end-to-end encrypted messaging are considered.

We have divided our recommendations into four categories that reflect different functions in Meta, with the goal of enabling Meta to more easily put our recommendations into action. However, many recommendations are relevant for multiple categories:



- **Product**—Recommendations about specific products (i.e., Messenger, Instagram DMs) and features, such as reporting, account linking, and discoverability
- **Process**—Recommendations for how Meta can detect and address human rights risks, such as user reporting and behavioral signals
- **Product Policy**—Recommendations for product policy across products, such as the Community Standards
- **Public Policy**—Recommendations for how Meta should engage with key external stakeholders, such as law enforcement and civil society

These recommendations are intended to shape how Meta should meet its responsibility to address the potential adverse human rights impacts arising from the expansion of end-to-end encryption, including how to exercise and increase its leverage to address those impacts.

It is important to note that many of the adverse human rights impacts associated with end-to-end encrypted messaging are system-wide and whole of society issues that exist beyond (and are often independent of) end-to-end encryption—such as sexual exploitation of children, human trafficking,



and terrorism and violent extremism. Governments are best placed to comprehensively address these kinds of issues, and indeed the UNGPs are clear that part of the state duty to protect human rights includes protecting their citizens from human rights abuses by third parties.

However, the UNGPs are also clear that companies have a responsibility to address any adverse

human rights impacts with which they are involved, including via collaboration with others. For this reason, many of BSR's recommendations are intended to help enable Meta to contribute to this ecosystem, including access to remedy—for example, by supporting entities that help victims of harm access justice and remediation services, or through improved user reporting channels.

## Address Risks and Maximize Benefits of End-to-End-Encryption



### Product

#### Recommendations about specific products and features:

- User reporting
- UX and user testing
- User education
- Languages
- Friction
- Opt-in account linking



### Process

#### Recommendations for how Meta can detect and mitigate human rights risk:

- Harm prevention strategies, such as metadata analysis
- ML techniques for proactive detection
- Holistic child rights strategy
- Investigate client-side scanning techniques
- Assess impacts of cross-app communications



### Product Policy

#### Recommendations for product policy changes:

- Consistent privacy policies and improved transparency
- Define content standards
- ML explainability
- Improved user reporting
- User appeals transparency
- Grievance mechanisms



### Public Policy

#### Recommendations for how Meta should engage external stakeholders:

- Advocate for end-to-end encryption
- Engage policy makers
- Engage stakeholders
- Collaborate with researchers
- Quantify harms
- Collaborate with industry
- Train law enforcement

## Areas for Future Assessment

In addition to end-to-end encryption, Meta also plans to make its messaging platforms capable of cross-app communication, enabling users on one platform to message users on another. This has important implications because each platform operates in a different context with different sets of user expectations. Just as with end-to-end encryption, there are several important decisions Meta will have to make about how it implements cross-app communication when that point is reached. Meta has announced that account linking will be optional,<sup>1</sup> and that users on one platform will be able to control whether they can be contacted by others on the other platforms,<sup>2</sup> but will users be able to control what kind of information about them is searchable by others? Will the Community Standards, currently applied to Messenger and Instagram DMs, apply to WhatsApp too? If not, what about messages sent between platforms—for example, from Messenger to WhatsApp? While this is not a HRIA of messaging cross-app communication, many decisions made about cross-app communication

will affect the human rights impacts of end-to-end encrypted messaging, and are therefore considered where relevant in this assessment.

There are also a variety of potential technical mitigation measures to address the human rights risks associated with end-to-end encryption that are not technically feasible today or have not yet been proposed. One example of this is homomorphic encryption, an approach to scanning content in its encrypted form that has been proposed as a mechanism to enable Meta to detect certain kinds of harmful and illegal content in end-to-end encrypted messaging. We discuss homomorphic encryption and other proposed content scanning solutions throughout this assessment, and particularly in Sections 10 and 11. However, because it is a nascent solution that is not yet technically feasible at scale in an end-to-end encrypted messaging context, our analysis and conclusions about it are inherently speculative. Further human rights assessment should be conducted if and when such solutions are technically feasible and can be more concretely explored.

<sup>1</sup> <https://thenextweb.com/news/whatsapp-opt-in-messenger-facebook-integration>.

<sup>2</sup> <https://www.facebook.com/help/messenger-app/2258699540867663>.

## 12.1 Product Recommendations

### User Reporting and User Interfaces

#### 1 RECOMMENDATION

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

Categories and labels for problematic content available via user reporting should be more consistent. At the time of writing, when users decide to report content or accounts, they are provided with divergent options for categorizing abuses in Instagram DMs and Messenger. Furthermore, it is only possible to report accounts or groups in WhatsApp. With the expansion of end-to-end encryption and cross-communication among messaging services, this could confuse users and discourage them from reporting abusive content or accounts.

At a minimum, Meta should present users with the same categories of abuse types to report across all messaging platforms. If Meta decides to have different Community Standards for each platform, it should also make the differences in reporting across platforms clear to users.

#### EXPLANATION

In shifting to end-to-end encrypted messaging services, Meta will need to rely heavily on user reporting to detect harmful content and abusive accounts. Social media researchers have shown that users have different habits in reporting abusive and problematic content, and many different factors might encourage or discourage them to report content and accounts.

This recommendation is based on BSR's conversation with human rights practitioners, civil society organizations, and academics who believe current methods of reporting, although necessary, are not an effective way to moderate abusive content and accounts. This is especially true for children, victims of trafficking, and people with lower levels of literacy and digital literacy. A recent report by the Center for Democracy & Technology (CDT) also concluded user reporting was a key method of content moderation in an end-to-end encrypted environment.<sup>3</sup>

Principle 29 of the UNGPs states, "Business enterprises should establish or participate in effective operational-level grievance mechanisms for individuals and communities who may be adversely impacted."

Principle 31 of the UNGPs states that non-judicial grievance mechanisms should be legitimate, accessible, predictable, equitable, transparent, rights-compatible, and a source of continuous learning. Of particular relevance to this recommendation is predictability: "providing a clear and known procedure with an indicative time frame for each stage, and clarity on the types of process and outcome available and means of monitoring implementation."

#### 2 RECOMMENDATION

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

<sup>3</sup> <https://cdt.org/insights/outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems/>.

Meta's Internationalization team should work with UX / UI research groups to expand their role beyond word-by-word translation of platform settings and policies to more tailored translations and information presentations for different contexts.

The two teams should also explore ways to customize user reporting interfaces based on different context and accessibility needs, especially for younger and differently abled children, and people with lower levels of digital literacy.

### EXPLANATION

User reporting is highly dependent on users' level of digital literacy, age, culture, language, and the context of reporting.<sup>4</sup> As such, a one-size-fits-all approach to user reporting may be insufficient for the range of users across Meta's messaging services.

Principle 31 of the UNGPs states that non-judicial grievance mechanisms should be legitimate, accessible, predictable, equitable, transparent, rights-compatible, and a source of continuous learning. Of particular relevance to this recommendation is accessibility: "being known to all stakeholder groups for whose use they are intended, and providing adequate assistance for those who may face particular barriers to access."

## 3 RECOMMENDATION

**Meta's UX/UI Research group should conduct participatory and co-design workshops to test user reporting features with children.**

The purpose of these workshops would be to identify how to protect children from unsolicited interactions, which might lead to grooming or trafficking.

These would take place with educators, child rights organizations, and children of different ages and genders and in different geographical regions to test different design features for how to navigate and report abusive accounts and content, pop-ups, and other account restriction features on all three messaging platforms.

### EXPLANATION

Children are highly vulnerable to exploitation, abuse, and other human rights harms via messaging services. In addition, they are often less likely to know how to respond to unsolicited interactions, and less likely to report content or other users in the case of exploitation or abuse. To address the increased risks for children, additional features or restrictions may be necessary.

This recommendation addresses the reality that many children below the age of 13 use private messaging services, even though terms of service may not allow them to.

Principle 18 of the UNGPs states that, to assess human rights impacts, companies should undertake "meaningful consultation with potentially affected groups."

The UNGPs state that companies should pay "particular attention to the rights and needs of, as well as the challenges faced by, individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized."

Principle 31 of the UNGPs states that operational-level mechanisms should be "based on engagement and dialogue: consulting the stakeholder groups for whose use they are intended on their design and performance, and focusing on dialogue as the means to address and resolve grievances."

<sup>4</sup> <https://doi.org/10.1177/1461444814543163>.

## 4 RECOMMENDATION

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

These metrics should be communicated in an accessible format in Meta’s transparency reporting portal. This assurance is especially important for groups who are most concerned about human rights risks of end-to-end encrypted messaging, including child rights groups, anti-human trafficking organizations, and counterterrorism and law enforcement agencies. Since Messenger and Instagram DMs are not end-to-end encrypted at the time of writing, there is an opportunity for Meta to develop and test different measurement techniques to explore the reliability of user reporting and other non-content-based mitigation techniques.

#### EXPLANATION

Principle 20 of the UNGPs states, “In order to verify whether adverse human rights impacts are being addressed, business enterprises should track the effectiveness of their response. Tracking should: (a) Be based on appropriate qualitative and quantitative indicators; (b) Draw on feedback from both internal and external sources, including affected stakeholders.”

Principle 21 of the UNGPs states, “In order to account for how they address their human rights impacts, business enterprises should be prepared to communicate this externally, particularly when concerns are raised by or on behalf of affected stakeholders.” Communications should “be of a form and frequency that reflect an enterprise’s human rights impacts and that are accessible to its intended audiences.”

Principle 31 of the UNGPs states that non-judicial grievance mechanisms should be a source of continuous learning, drawing on relevant measures to identify lessons for improving the mechanism and preventing future grievances and harms.

## 5 RECOMMENDATION

### **Explore and define how to verify the authenticity of users’ reports.<sup>5</sup>**

Because Meta will be relying heavily on user reporting to address abuse and problematic content in end-to-end encrypted messaging, it is important for the company to be able to verify the authenticity of reports without weakening end-to-end encryption.

A technique known as “message franking” is currently used for Messenger to ensure that a report originated from the reporter’s device. Moving forward, Meta should consider the pros and cons of applying message franking for WhatsApp and Instagram DMs. This is especially important in the context of cross-app communication of the three messaging platforms.

#### EXPLANATION

Principle 20 of the UNGPs states, “In order to verify whether adverse human rights impacts are being addressed, business enterprises should track the effectiveness of their response. Tracking should: (a) Be based on appropriate qualitative and quantitative indicators; (b) Draw on feedback from both internal and external sources, including affected stakeholders.”

Principle 31 of the UNGPs states that non-judicial grievance mechanisms should be legitimate, enabling trust from the stakeholder groups for whose use they are intended, and being accountable for the fair conduct of grievance processes.”

<sup>5</sup> <https://eprint.iacr.org/2017/664.pdf>.

## 6 RECOMMENDATION

### **Invest in ensuring that users who have violated platform policies cannot return.**

Once a user has been kicked off a Meta platform for egregious and confirmed violations of Community Standards (i.e., through sharing CSAM, inappropriate interactions with minors, etc.) they should not be able to return unless an appeal has found them to be innocent of the charge. Meta should invest in detection tools to prevent abusers from returning both overtly or covertly, for example by using other aliases and phone numbers.

### EXPLANATION

This recommendation is designed to help keep vulnerable groups, especially children, safe on Meta's messaging platforms.

Principle 19 of the UNGPs states, "In order to prevent and mitigate adverse human rights impacts, business enterprises should integrate the findings from their impact assessments across relevant internal functions and processes, and take appropriate action."

Principle 19 of the UNGPs further states, "If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it."

Principle 25 of the UNGPs references "guarantees of non-repetition" as a type of remedy.

## User Education Features

## 7 RECOMMENDATION

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

Meta's recent collaboration with the World Health Organization (WHO) to create the "WHO Health Alert" is an example of such in-app services, and can be applied in different contexts.<sup>6</sup> For instance, by:

- Embedding rapid response / hotline services within the messaging platforms, via automated services (such as rapid response bots) or connecting to human support and service providers. This is especially important in cases of sex trafficking because victims often reach out to support and service providers by initiating a conversation on messaging platforms.
- Embedding in-app reliable fact-checking services such as fact-checking bots in order to minimize the risk of virality of mis / disinformation content.

### EXPLANATION

In addition to providing users a reporting option, Meta can develop additional in-app features that provide accurate information or resources. This is especially important for vulnerable users who have been harmed or threatened by harm in some way.

Such support features are important for helping users in the moment they need it. Comparatively, reporting typically takes time and only results in penalties for the violating user, not support for the reporting user.

Principle 19 of the UNGPs states, "In order to prevent and mitigate adverse human rights impacts, business

<sup>6</sup> <https://www.whatsapp.com/coronavirus/who>.

enterprises should integrate the findings from their impact assessments across relevant internal functions and processes, and take appropriate action.”

The UNGPs state that companies should pay “particular attention to the rights and needs of, as well as the challenges faced by, individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized.”

## Group Size and Message Forwarding

### 8 RECOMMENDATION

**Assess options for adding “friction” for contacting groups and strangers on its messaging platforms in order to minimize unsolicited interactions, virality of harmful mis/disinformation and hate speech, inauthentic behavior, and other actions that may lead to negative human rights impacts.**

This can be achieved by:

- Continuing to experiment with limiting users’ ability to message people they don’t know.
- Continuing to experiment with limiting the size of group messages.
- Continuing to experiment with limiting “forwardability” in terms of number of forwarded messages, number of contacts to whom a message is forwarded, and types of contacts (whether they are in user’s contact list or not).
- For group messages, it is important for Meta to differentiate between the features available to group admins vs. group members. This is especially important in the context of viral mis / disinformation and in-group harassment, in which group admins have power to moderate the conversation and members to some extent.

In developing such experiments, Meta should differentiate between the types of friction applied in public groups, private groups, and one-to-one chat because the nature of content creation, circulation, and moderation are very different and each might impact users’ human rights differently.

### EXPLANATION

This recommendation is informed by BSR’s research on the concerns around “coordinated inauthentic behavior” and virality of mis / disinformation and other harmful content on end-to-end encrypted messaging platforms. WhatsApp’s experiments with limiting the number of forwards and labeling forwarded messages are promising steps toward minimizing virality of problematic content.

Principle 19 of the UNGPs states, “In order to prevent and mitigate adverse human rights impacts, business enterprises should integrate the findings from their impact assessments across relevant internal functions and processes, and take appropriate action.”

Principle 19 of the UNGPs further states, “If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it.”

## Messenger Kids

### 9 RECOMMENDATION

**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.**

Currently, Messenger Kids offers significant parental controls via account linking and a robust dashboard. Without these protections, Messenger Kids risks becoming a space that is ripe for exploitation by bad actors seeking to target children. However, if these safeguards can be retained in the context of end-to-end encryption, then implementing end-to-end encryption in Messenger Kids would bring significant rights protection (such as privacy and freedom of expression) to children.

BSR notes that this does not mitigate all potential harms to children, because children may still use Messenger, Instagram DMs, and WhatsApp.

BSR also notes that, at the time of writing, there are no plans to bring end-to-end encryption to Messenger Kids.

### EXPLANATION

As highly vulnerable users, children should receive greater protections from the risks of end-to-end encrypted messaging. Though children may use Meta's other messaging services, Messenger Kids represents a clear space where greater protections can be implemented without adverse impacts to other rightsholders who would benefit from the privacy protections of end-to-end encryption.

Principle 19 of the UNGPs states, "In order to prevent and mitigate adverse human rights impacts, business enterprises should integrate the findings from their impact assessments across relevant internal functions and processes, and take appropriate action."

Principle 19 of the UNGPs further states, "If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it."

## Registration and Account Linking

### 10 RECOMMENDATION

**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.**

Account linking leads to discoverability of users across all Meta platforms; the implications of discoverability need to be carefully considered and users who have opted out of account linking should not be discoverable across platforms. For example, many of the integrity mitigations currently used on WhatsApp would be undermined by discoverability.

In addition, Meta should be transparent with users about the privacy concerns of discoverability in the context of registration and account linking. Users should be prompted to provide informed consent before account linking takes place. If a user does consent, they should still be provided with the option to opt out at a later date.



## EXPLANATION

Currently, each messaging platform has its own registration method: WhatsApp users sign-up with their phone number; to use Messenger, users have to create a Meta account with their real name, and to use Instagram DMs users must create an Instagram account which doesn't need users' phone number, email, or real name.<sup>7</sup>

This inconsistency in the registration process and account linking might create privacy concerns for the most vulnerable groups, who may only feel private and safe online if they are anonymous and undiscoverable.

Principle 19 of the UNGPs states, "In order to prevent and mitigate adverse human rights impacts, business enterprises should integrate the findings from their impact assessments across relevant internal functions and processes, and take appropriate action."

## 12.2 Process Recommendations

### Proactive Detection of Abusive Accounts and Content

#### 11 RECOMMENDATION

**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.** Many of these methods have already proved to be quite effective at preventing certain types of human rights harm, such as child grooming.

Meta should continue investigating such methods, and be more public about their utility and effectiveness. This is important not only to share lessons learned with all stakeholders in the pursuit of preventing online harm in general, but also to prove that end-to-end encryption does not preclude harm prevention and reduction.

## EXPLANATION

Although policy actors tend to focus disproportionately on identifying and removing content, there is an important role for Meta to play in preventing human rights harm from occurring in the first place. Whereas content detection and removal is made less feasible by end-to-end encryption, harm prevention need not be.

Principle 19 of the UNGPs states, "If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it."

<sup>7</sup> <https://help.instagram.com/494561080557017>.

## 12 RECOMMENDATION

**During the design and development of machine learning-enabled 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.**

Meta's “Fairness in Machine Learning” team should work with Meta product teams in the design and development of such ML-based systems, and consider human rights factors during the process. This might include considering the interests of / impacts on different vulnerable groups, and engaging with external human rights expertise.

It is also important to consider adequate remedy in these contexts, as classifier-based methods will inevitably make errors that result in users being erroneously suspended or removed from platforms. Ensuring clear and easy access to effective appeals mechanisms is therefore important.

The need for transparency and interpretability of such systems is also especially important in the context of account linking and cross-app communication of messaging platforms.

### EXPLANATION

ML approaches can be helpful in proactively detecting risk of harm, but should be designed and deployed with care. Human rights researchers and practitioners have raised concerns regarding adverse human rights impacts of ML systems, including on the right to privacy, the rights to equality and nondiscrimination, the right to freedom of expression, the right to an effective remedy, and more.<sup>8</sup> These concerns can typically be addressed through careful consideration of the fairness of training datasets and algorithms, human-in-the-loop models, and assessments.

During BSR's external interview process, several interviewees raised concerns around the efficacy of relying solely on ML-enabled techniques for detecting abusive content. They believe that, at the moment, such systems are not advanced enough to detect abusive content, especially in highly contextual situations, including but not limited to child grooming, detecting terrorist activities, illicit good sales, and human trafficking. Some interviewees raised concerns about both vagueness and privacy implications of the practice of collecting “behavioral signals.” They believe relying on such signals to train and optimize ML-enabled systems might lead to a high number of false positive and false negative cases, especially for the most vulnerable groups whose local languages and user behavior are often rarely represented (or, conversely, overrepresented) during training such systems.

As recommended in the latest pilot study by the Ranking Digital Rights research group, it is important for Meta to develop an “Algorithmic Systems Use Policy” to ensure transparency and accountability of using algorithmic decision-making systems.<sup>9</sup> The details of such a policy is described in the Product Policy section of the recommendations.

<sup>8</sup> A few examples include: Report by the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/73/348, 2018, <https://freedex.org/wp-content/blogs.dir/2015/files/2018/10/AI-and-FOE-GA.pdf>; CDT, Mixed Messages? The Limits of Automated Social Media Content Analysis, November 28, 2017, <https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/>; and The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems, 2018, [https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration\\_ENG\\_08-2018.pdf](https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf)

<sup>9</sup> <https://rankingdigitalrights.org/wp-content/uploads/2020/03/pilot-report-2020.pdf>.

## Proactive Detection of Child Sexual Abuse Material

### 13 RECOMMENDATION

**Create a child rights strategy for end-to-end encrypted messaging services that brings together all the elements needed to address risks to child rights holistically.**

This should include, but not be limited to:

- Accessible user reporting features
- User education
- Metadata analysis
- Use of behavioral signals
- Further investigating scalable and client-side scanning techniques for CSAM that maintain the cryptographic integrity of end-to-end encryption
- Law enforcement training and partnerships
- Civil society partnerships
- The development of metrics to quantify the scope of CSAM and corresponding harm, among others



### EXPLANATION

The challenges in addressing the child rights risks associated with end-to-end encrypted messaging are complex and multifaceted, and a systemic approach that is both reactive and preventive is needed. A holistic child rights strategy is key for Meta to be able to appropriately address the risks to child rights as they evolve over time and as the potential mitigations measures grow and change.

Principle 19 of the UNGPs states, “If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it.”

Principle 17(c) of the UNGPs states that human rights due diligence “should be ongoing, recognizing that the human rights risks may change over time as the business enterprise’s operations and operating context evolve.”

## 14 RECOMMENDATION

**(Part 1) 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.**

Some security experts have proposed homomorphic encryption as an approach that would in theory allow Meta to detect CSAM in end-to-end encrypted messaging while maintaining cryptographic integrity, though a debate remains over whether these methods would still undermine the end-to-end encryption by violating the principle that all information about the content of a message should be known only to the sender and intended recipients (see Section 5.3).<sup>10</sup>

However, using computing power available today, homomorphic encryption is not currently technically feasible at the scale of Meta (which has messaging services used by billions), and therefore cannot yet be deployed.

Nevertheless, BSR believes there is value to be gained from researching homomorphic encryption (and other potential cryptographic integrity preserving methods) further, and Meta should devote resources to investigating them, including both their technical feasibility and the human rights risks that may arise from them (e.g., the “slippery slope” risk). It will be important for Meta to be an active and informed participant in dialogue about client-side scanning methods, given that they will likely become more technically feasible as research and the computational power of mobile devices improve.

Further, Meta can improve its external engagement on this specific issue, collaborate with researchers and child protection organizations in both research and testing, and contribute findings to the tech industry so that other companies can learn.

As discussed previously in this report, client-side scanning does not solve the problem of live or new / unknown CSAM. Unfortunately, there are currently no known methods of detecting new / unknown images or live child exploitation via video chat that do not involve weakening or breaking end-to-end encryption.

### EXPLANATION

This recommendation is based on counterbalancing the right to privacy for all users and protection of children from exploitation discussed earlier in this report. The counterbalancing exercise revealed that while end-to-end encryption is the only way to truly protect the privacy of communications today, and thus should not be weakened or banned, restricting the right of children to be protected from sexual abuse and exploitation in favor of protecting user privacy by deploying end-to-end encryption without any form of content-based CSAM detection may not be necessary or proportionate if there are approaches, such as homomorphic encryption, that could maintain all of the privacy protections of end-to-end encryption while enabling improved protection of child rights. See the following recommendation for caveats to this statement.

Principle 19 of the UNGPs states that “If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it.”

<sup>10</sup> See: Jonathan Mayer, “Content Moderation for End-to-End Encrypted Messaging” Princeton University, October 6, 2019, [https://www.cs.princeton.edu/~jrmayer/papers/Content\\_Moderation\\_for\\_End-to-End\\_Encrypted\\_Messaging.pdf](https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf); Priyanka Singh and Hany Farid, “Robust Homomorphic Image Hashing,” Computer Vision Foundation Workshop, [http://openaccess.thecvf.com/content\\_CVPRW\\_2019/papers/Media%20Forensics/Singh\\_Robust\\_Homomorphic\\_Image\\_Hashing\\_CVPRW\\_2019\\_paper.pdf](http://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Singh_Robust_Homomorphic_Image_Hashing_CVPRW_2019_paper.pdf); Hany Farid, “Opinion: Facebook’s Encryption Makes it Harder to Detect Child Abuse,” Berkeley School of Information, October 25, 2019, <https://www.ischool.berkeley.edu/news/2019/opinion-facebook-encryption-makes-it-harder-detect-child-abuse>.

## 15 RECOMMENDATION

**(Part 2) 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 (e.g., privacy, freedom of expression) and a conclusion that those impacts can be adequately addressed.**

Meta should conduct human rights due diligence on any potential client-side solutions to account for design decisions, technical factors, and the legal, political, and regulatory context.

If the implementation of client-side scanning solely to detect CSAM—a legitimate aim—would result in adverse impacts on privacy, freedom of expression, and other rights in a way that is inconsistent with the principles of necessity, proportionality, and nondiscrimination, then client-side scanning should not be pursued, even if technically feasible. In this case, the adverse human rights impacts enabled by client-side scanning for CSAM detection would constitute a disproportionate restriction on the rights of all users, and alternative methods of addressing child exploitation risks should be pursued instead.

The slippery slope risk in today’s legal, political, and regulatory context is real, and for this reason, BSR notes that client-side scanning capable of detecting CSAM while maintaining the cryptographic integrity of end-to-end encryption may never be implemented or may only be suitable for some products but not others.

BSR reaches this conclusion based on human rights factors, rather than a point of view about whether a narrow definition of end-to-end encryption (focused on cryptographic integrity and technical process) or broad definition (focused on who knows about the content of a message) should be adopted. This conclusion—that there may always be human rights-based barriers to client-side scanning—reinforces the need for alternative methods of CSAM detection (such as more effective reporting and metadata analysis) to be researched, developed, and deployed.

If Meta decides to implement any methods that involve client-side scanning for CSAM in the future, it should ensure that the technique is compatible with devices that have different storage and computation capacity. If device capability is a near-term challenge, Meta should seek to run client-side scanning on all capable devices—i.e., this need not be an all-or-nothing approach.

### EXPLANATION

Principle 17 of the UNGPs states that human rights due diligence should be initiated as early as possible in the development of a new activity.

Principle 18 of the UNGPs states that human rights assessment should take place prior to a new business activity.

## Cross-App Communication

## 16 RECOMMENDATION

**Conduct human rights due diligence on cross-app communication.**

This was *not* an HRIA of the cross-app communication of Meta’s messaging platforms. Although we explored many aspects of cross-app communication in this report, this cannot be considered a comprehensive assessment of all the potential adverse human rights impacts of cross-app communication. We recommend Meta conduct human rights due diligence to further explore the human rights implications of the decisions it makes about cross-app communication.

## EXPLANATION

Principle 15 of the UNGPs states, “In order to meet their responsibility to respect human rights, business enterprises should have in place policies and processes appropriate to their size and circumstances, including a human rights due diligence process to identify, prevent, mitigate, and account for how they address their impacts on human rights.”

Principle 18 of the UNGPs states that “assessments of human rights impacts should be undertaken ... prior to major decisions or changes in the operation,” including market entry, product launch, policy change, or wider changes to the business.

## 12.3 Product Policy Recommendations

### Privacy Policy and Informed Consent

#### 17 RECOMMENDATION

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

Cross-app communication of Meta’s three messaging platforms (especially in the case of account linking and discoverability) could make it impossible to maintain separate privacy policies across platforms.

However, inconsistency between private and public features of each platform also confuses users about what data each and all platforms collect, share, and retain (e.g., at the moment Instagram DMs and the Instagram App share the same privacy policy).<sup>11</sup> Therefore, in its privacy policies, Meta should:

- Be transparent about the types of metadata collected from users and how cross-app communication and account linking might affect it.
- Describe “behavioral signals” in a way that strikes a balance between being informative to users and avoiding the potential for bad actors to game the system. This includes a description of how these signals might be collected, shared, and used to train ML systems that would help Meta to detect problematic content and accounts.
- Clearly describe the purpose of “behavioral signals.” Meta must ensure that these signals, if meant only to detect and mitigate abusive and problematic content, are not used and shared for other business-related practices such as ad targeting.
- Be clear whether a user’s activities on Meta’s public platforms might affect Meta’s decisions about the user’s private messaging accounts. This is especially important in the case of Instagram and Facebook because users of these platforms engage in both public activities and private activities, and users have the right to understand what data are collected and how that data are used.

## EXPLANATION

Indicator P7 in the Ranking Digital Rights Index states, “The company should clearly disclose to users what options they have to control the company’s collection, retention and use of their user information.”

<sup>11</sup> <https://help.instagram.com/519522125107875>.

Principle 16 of the UNGPs states that commitment to human rights should be “reflected in operational policies and procedures necessary to embed it throughout the business enterprise.”

## Community Standards and Terms of Service

### 18 RECOMMENDATION

**Apply a minimum level of consistency in community standards across all messaging platforms to facilitate improved user reporting.**

User reporting will be key to detecting and mitigating abuse in an end-to-end encrypted environment. To facilitate user reporting, there must be standards that support reports—and this means that all messaging platforms should have some level of uniformity for the categories of abuse types that can be reported and the terminology used.

However, the question remains—should Meta have one set of community standards that apply to all messaging platforms? Or should Meta maintain different sets of standards? If there is one set of standards, should they be extensive like the Facebook Community Standards or minimalist like the current WhatsApp standards? Unfortunately, there is no clear answer to this.

On one hand, having the same set of standards would make content moderation decisions and user reporting easier, and would improve access to remedy. It would also avoid potential standards conflicts between users messaging across platforms.

On the other hand, creating one-size-fits-all standards might confuse users, who consent to a different kind of service when joining WhatsApp vs. Facebook vs. Instagram DMs and thus have different expectations. Having one set of standards could also be tricky because the different messaging platforms have different features that necessitate specific policies, such as peer-to-peer payments.

It will also be important to consider the enforceability of community standards in end-to-end encrypted messaging. For example, the Facebook community standards are quite broad and comprehensive, and often require significant context to enforce. Hate speech, incitement to violence, and bullying and harassment are examples of this.

#### EXPLANATION

WhatsApp does not currently have community standards because originally the app was intended to be used as an alternative to SMS-based private messaging services. However, while WhatsApp is still a private messaging app, the existence of large groups has changed the nature of the app. In the near future, with the cross-app communication of all three messaging platforms, the ways in which users use Meta’s messaging platform might change significantly.

Principle 16 of the UNGPs states that commitment to human rights should be “reflected in operational policies and procedures necessary to embed it throughout the business enterprise.”

### 19 RECOMMENDATION

**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.**

The Oversight Board exists in part to provide guidance to Meta on difficult product policy decisions that are a matter of public debate. The question of whether and what community standards should apply across end-to-

end encrypted messaging platforms is ultimately about the broader emerging debate about content standards and moderation in private messaging services, and is a judgement call that would benefit from Oversight Board review.

#### EXPLANATION

Principle 16 of the UNGPs states that commitment to human rights should be “reflected in operational policies and procedures necessary to embed it throughout the business enterprise.”

While this question is not a content-specific case within the current scope of the Oversight Board, it is the type of product policy issue on which the Board could provide valuable feedback and insight, and which holds considerable significance for future content decisions that may come before the Oversight Board.

### 20 RECOMMENDATION

**In cases where separate standards conflict, Meta should always apply the stricter standard.**

If Meta were to keep separate standards for each platform, conflicts between users messaging across platforms would surely occur. Applying the stricter standard is necessary to maximize protections for users and to meet user’s expectations for the platform. This recommendation is only applicable in the context of cross-app communication and does not apply to private messaging platforms used in isolation.

#### EXPLANATION

Principle 16 of the UNGPs states that commitment to human rights should be “reflected in operational policies and procedures necessary to embed it throughout the business enterprise.”

Principle 31 of the UNGPs lays out effectiveness criteria for non-judicial grievance mechanisms, including two criteria—predictability and transparency—that are important for setting rightsholders’ expectations. In this situation, Meta should also be clear and predictable regarding which standards apply to users’ content.

## ML Classifiers for Abuse Detection

### 21 RECOMMENDATION

**Develop publicly available, accessible, and understandable policy documents to disclose its use of classifiers in detecting, flagging, and moderating accounts and content on its messaging platforms.**

It should be clear for users what kinds of variables are used to make decisions about their accounts and content and how much control they have to opt in or out from such forms of algorithmic decision-making.

BSR recognizes that some public disclosure may not be possible or advisable for various reasons, such as trade secrets, or to prevent the possibility of malicious actors abusing and gaming the systems. To the extent possible, Meta should share information about its use of classifiers in messaging platforms.

#### EXPLANATION

This recommendation is derived from the Ranking Digital Rights pilot study on the importance of transparency and accountability for using algorithmic decision-making systems.

In addition, Principle 21 of the UNGPs states that “in order to account for how they address their human rights impacts, business enterprises should be prepared to communicate this externally, particularly when concerns



are raised by or on behalf of affected stakeholders.” Communications should “be of a form and frequency that reflect an enterprise’s human rights impacts and that are accessible to its intended audiences.”

According to GNI Principle 6.2, company “participants will be held accountable through a system of (a) transparency with the public and (b) independent assessment and evaluation of the implementation of [GNI] Principles.”

Principle 16 of the UNGPs states that commitment to human rights should be “reflected in operational policies and procedures necessary to embed it throughout the business enterprise.”

## 22 RECOMMENDATION

### **Examine whether and how classifiers for detecting, flagging, and moderating accounts and content on its messaging platform could result in discrimination.**

Meta’s Machine Learning Fairness team should develop methods to examine how these systems might violate users’ right to equality and nondiscrimination, especially for vulnerable populations along lines of gender, race, ethnicity, language, or other socioeconomic groups.

#### EXPLANATION

The introduction of the UNGPs states, “These Guiding Principles should be implemented in a nondiscriminatory manner, with particular attention to the rights and needs of, as well as the challenges faced by, individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized, and with due regard to the different risks that may be faced by women and men.”

## 23 RECOMMENDATION

### **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.**

This would ensure that ML systems are audited against criteria such as fairness and the right to nondiscrimination, especially with respect to vulnerable groups.

Auditing processes might be done in various forms, such as bias and fairness red-teaming and scenario-simulations (internally) or fairness bounty programs (externally).<sup>12</sup>

There is potential to use human rights impact assessment methodology in developing scenarios to assess how Meta’s algorithmic systems used in end-to-end encrypted messaging apps might affect the most vulnerable groups of users. These scenarios should be informed by Meta’s ethnographic research on user behavior and human rights concerns in different regions of the world, and by active communication with civil society organizations.

#### EXPLANATION

This recommendation is derived from a Meta research team’s recent paper on the possibility of simulating users’ behavior online.<sup>13</sup>

The introduction to the UNGPs states, “These Guiding Principles should be implemented in a nondiscriminatory manner, with particular attention to the rights and needs of, as well as the challenges faced by, individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized, and with due regard to the different risks that may be faced by women and men.”

<sup>12</sup> <https://arxiv.org/abs/2004.07213>.

<sup>13</sup> <https://research.fb.com/wp-content/uploads/2020/04/WES-Agent-based-User-Interaction-Simulation-on-Real-Infrastructure.pdf>.

## Transparency Reporting

### 24 RECOMMENDATION

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

The Meta transparency reporting portal should distinguish between content on open platforms and content in private messaging services, and disaggregate data in a more meaningful way for readers, especially for civil society groups, who seek to understand how Meta enforces its Community Standards and terms of service to protect the most vulnerable groups.

Currently, Meta’s Community Standards Enforcement Report does not disaggregate data based on geographical region or other possible factors such as gender or age.<sup>14</sup> Wherever possible, via consultation with civil society groups, Meta should disaggregate reported data in its transparency report. For example, in the case of sexual abuse and exploitation of children, data can be further disaggregated based on gender, age, and forms of exploitation.

#### EXPLANATION

The recommendation on the need for more disaggregated data is derived from the UN Convention on the Rights of the Child guidelines, which highlights the importance of implementing “a disaggregated approach to data, addressing how these offenses [including sale of children, child prostitution, and child pornography] affect different groups of children. At a minimum, data should be disaggregated by sex, age, and form of exploitation.”<sup>15</sup>

This recommendation is also derived from the UNGPs Gender Framework, which states that “Business enterprises should track the effectiveness of their responses by using sex-disaggregated data, collected in line with a human rights-based approach, and outcome indicators developed in consultation with affected women, women’s organizations, and gender experts.”

Civil society actors believe that further disaggregation of categories of harassment and abuse is needed to understand the scale of abuse and harassment issues on Meta and the company’s efficacy in applying its Community Standards. In a study on violence and abuse against women on Twitter, Amnesty International also noted the importance of providing disaggregated data by category of abuse in transparency reporting.<sup>16</sup>

Principle 21 of the UNGPs states, “In order to account for how they address their human rights impacts, business enterprises should be prepared to communicate this externally, particularly when concerns are raised by or on behalf of affected stakeholders.” Communications should “be of a form and frequency that reflect an enterprise’s human rights impacts and that are accessible to its intended audiences.”

According to GNI Principle 6.2, company “participants will be held accountable through a system of (a) transparency with the public and (b) independent assessment and evaluation of the implementation of [GNI] Principles.”

<sup>14</sup> <https://transparency.facebook.com/community-standards-enforcement#hate-speech>.

<sup>15</sup> [https://www.ohchr.org/Documents/HRBodies/CRC/CRC.C.156\\_OPSC%20Guidelines.pdf](https://www.ohchr.org/Documents/HRBodies/CRC/CRC.C.156_OPSC%20Guidelines.pdf),

<sup>16</sup> <https://decoders.amnesty.org/projects/troll-patrol/findings>.

## Government Requests for Information

### 25 RECOMMENDATION

**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.**

For example, governments might request increasing amounts of metadata and public profile information in the absence of access to message content.

Meta should ensure that its approach to governments and third-party data sharing is in line with respecting users' privacy and freedom of expression rights, as expressed in the Global Network Initiative Principles and Implementation Guidelines.

If governments begin requesting new types of information or Meta notices new trends, these should be communicated in Meta's Transparency Report.

#### EXPLANATION

The GNI principles state, "Participating companies will employ protections with respect to personal information in all countries where they operate in order to work to protect the privacy rights of users. Participating companies will respect and work to protect the privacy rights of users when confronted with government demands, laws, or regulations that compromise privacy in a manner inconsistent with internationally recognized laws and standards."

## Enforcement Actions, User Appeals, and Effective Remedy

### 26 RECOMMENDATION

**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.**

For example, prior to suspending an account, Meta could warn the user that they have detected potential problematic behavior, and offer the user the chance to appeal or dispute. Meta could also seek to more proactively discourage abusive behavior via prompts and behavioral nudges.

#### EXPLANATION

Meta's decision to use "behavioral signals" to inform decisions about accounts and content especially affects its ability to explain the decision-making process. Meta products are a key vector for user expression of human rights and thus users have the right to understand why their accounts are suspended or removed and how they can appeal Meta's decision. This also applies to the user reporting mechanism.

Principle 22 of the UNGPs states, "Where business enterprises identify that they have caused or contributed to adverse impacts, they should provide for or cooperate in their remediation through legitimate processes."

According to Ranking Digital Rights' 2019 Ranking Digital Rights Corporate Accountability Index, a key recommendation to Meta was to "improve appeals mechanisms [by improving] its grievance and remediation mechanisms for users whose freedom of expression and privacy are violated by the company's policies and practices."<sup>17</sup>

The recommendation is also derived from BSR's research and conversation with civil society groups who

<sup>17</sup> <https://rankingdigitalrights.org/index2019/companies/facebook/index/>.

believe that the appeals process in cases of user reporting and account takedowns is too slow and difficult to navigate, especially for children and users with lower levels of literacy and digital literacy.

Principle 31 of the UNGPs states that non-judicial grievance mechanisms should be legitimate, accessible, predictable, equitable, transparent, rights-compatible, and a source of continuous learning.

## 27 RECOMMENDATION

### **Provide more information about how Meta’s appeals processes work on end-to-end encrypted messaging platforms.**

Unlike on its open platforms, there is no option for restoring lost content in messages. Users should be informed about what kinds of enforcement decisions they can appeal, such as suspensions or bans. In its Transparency Reporting portal Meta can disclose what remediation options users have on each end-to-end encrypted platform and how the appeals and remediation process takes place on those platforms. In addition, Meta can disclose the average time it takes for the company to make decisions on appeals and account restoration requests based on different categories and on each service.

#### EXPLANATION

Principle 31 of the UNGPs states that non-judicial grievance mechanisms should be legitimate, accessible, predictable, equitable, transparent, rights-compatible, and a source of continuous learning. Of particular relevance to this recommendation is transparency: “keeping parties to a grievance informed about its progress, and providing sufficient information about the mechanism’s performance to build confidence in its effectiveness and meet any public interest at stake.”

## 28 RECOMMENDATION

### **Increase the speed and capacity of reporting and appeals processes, especially for vulnerable groups.**

Reporting and appeals processes can stretch on for weeks, often resulting in harm because of the lost window of opportunity for restored accounts or content, particularly during major events or crisis periods. Meta should provide ways for users to track the status of their reports and provide a final decision about those reports in a more consistent, timely manner.

#### EXPLANATION

Principle 31 of the UNGPs states that non-judicial grievance mechanisms should be legitimate, accessible, predictable, equitable, transparent, rights-compatible, and a source of continuous learning. Of particular relevance to this recommendation is predictability: “providing a clear and known procedure with an indicative time frame for each stage, and clarity on the types of process and outcome available and means of monitoring implementation.”

## 29 RECOMMENDATION

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

Meta's appeals mechanisms are essentially operational grievance mechanisms, and the UNGPs effectiveness criteria set out the baseline requirements appeals mechanisms must achieve.

For example, the ideal reporting channel would be designed to meet the needs of billions of rightsholders who could be anywhere in the world, who may speak any language, and who have a wide range of different digital capabilities.

### EXPLANATION

Principle 31 of the UNGPs states that in order to ensure their effectiveness, non-judicial grievance mechanisms, both State-based and non-State-based, should be:

- (a) Legitimate: enabling trust from the stakeholder groups for whose use they are intended, and being accountable for the fair conduct of grievance processes;
- (b) Accessible: being known to all stakeholder groups for whose use they are intended, and providing adequate assistance for those who may face particular barriers to access;
- (c) Predictable: providing a clear and known procedure with an indicative time frame for each stage, and clarity on the types of process and outcome available and means of monitoring implementation;
- (d) Equitable: seeking to ensure that aggrieved parties have reasonable access to sources of information, advice, and expertise necessary to engage in a grievance process on fair, informed, and respectful terms;
- (e) Transparent: keeping parties to a grievance informed about its progress, and providing sufficient information about the mechanism's performance to build confidence in its effectiveness and meet any public interest at stake;
- (f) Rights-compatible: ensuring that outcomes and remedies accord with internationally recognized human rights;
- (g) A source of continuous learning: drawing on relevant measures to identify lessons for improving the mechanism and preventing future grievances and harms;

Operational-level mechanisms should also be: (h) Based on engagement and dialogue: consulting the stakeholder groups for whose use they are intended on their design and performance, and focusing on dialogue as the means to address and resolve grievances.

## Internal Governance of Meta

### 30 RECOMMENDATION

#### **Integrate human rights due diligence into its privacy review and data protection assessment procedures.**

Meta should also improve its regional knowledge and representation (especially in the countries from the Global South where the majority of Meta users reside) by diversifying its talent pipeline and closing the silos between different internal teams, including product policy, safety and security, partnerships, public policy, and research and engineering.

### EXPLANATION

This recommendation is intended to better integrate consideration of human rights into Meta's decision-making processes.

Principle 16 of the UNGPs states that commitment to human rights should be “reflected in operational policies and procedures necessary to embed it throughout the business enterprise.”

Principle 17(c) of the UNGPs states that human rights due diligence “should be ongoing, recognizing that the human rights risks may change over time as the business enterprise’s operations and operating context evolve.”

According to GNI 4.2.7 (f), companies should “update human rights impact assessments over time, such as when there are material changes to laws, regulations, markets, products, technologies, or services” and “develop internal processes and mechanisms for using the results of impact assessments to inform company policy and practice.”

## 12.4 Public Policy Recommendations

### Advocacy

#### 31 RECOMMENDATION

**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.**

BSR recognizes the challenges of advocacy efforts on end-to-end encryption, especially given the polarized debate between privacy and security. For this reason, Meta could also provide educational opportunities for policymakers, law enforcement agencies, and civil society organizations to learn more about end-to-end encryption from a technical perspective, and to openly discuss mitigation techniques that can be leveraged to minimize harms from the expansion of end-to-end encryption.

#### EXPLANATION

The GNI Principles state that “individually and collectively, participants will engage governments and international institutions to promote the rule of law and the adoption of laws, policies and practices that protect, respect and fulfil freedom of expression and privacy.”

Principle 19 of the UNGPs states that “If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it. And if it lacks leverage there may be ways for the enterprise to increase it. Leverage may be increased by, for example, offering capacity-building or other incentives to the related entity, or collaborating with other actors.”

#### 32 RECOMMENDATION

**Meta should engage policymakers about conflicting regulatory requirements that unnecessarily pit privacy rights against protecting users from broader harm.**

Efforts such as the new EU e-Privacy Directive, which prevents companies that operate messaging services from using metadata other than to facilitate the sending and receipt of messages, have an adverse impact on Meta’s ability to address the human rights risks identified in this assessment. While intended to protect the privacy rights of users, in practice this prevents Meta from using metadata to mitigate human rights risk and

prevent harm in end-to-end encrypted messaging, and conflicts with other legal requirements to identify and remove illegal content.

Meta should engage with policymakers about the existence and nature of these conflicts, and how metadata can be used in ways that are necessary and proportionate.

#### EXPLANATION

The GNI Principles state that “individually and collectively, participants will engage governments and international institutions to promote the rule of law and the adoption of laws, policies and practices that protect, respect and fulfil freedom of expression and privacy.”

Principle 19 of the UNGPs states that “If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it. And if it lacks leverage there may be ways for the enterprise to increase it. Leverage may be increased by, for example, offering capacity-building or other incentives to the related entity, or collaborating with other actors.”

### Partnerships and Collaboration

#### 33 RECOMMENDATION

**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.**

Because its messaging products are used by billions of people around the world, the adverse impacts associated with Meta’s expansion of end-to-end encryption have a significant global impact, and any decisions it makes related to the design and deployment of end-to-end encryption and mitigations it pursues have cascading effects. Meta’s actions could also be seen as standard setting for end-to-end encrypted messaging services more broadly. Thoughtful and deliberate collaboration with external stakeholders is therefore vital to Meta’s success in addressing human rights risks and pursuing human rights opportunities.

#### EXPLANATION

The GNI Principles state, “The development of collaborative strategies involving business, industry associations, civil society organizations, investors and academics will be critical to the achievement of these Principles.”

GNI Principle 5 states that “Participants will take a collaborative approach to problem solving and explore new ways in which the collective learning from multiple stakeholders can be used to advance freedom of expression and privacy.”

GNI Principles Implementation Guideline 4.2 states, “Application Guidance: Promoting rule of law reform could include rule of law training, capacity building with law-related institutions, taking public policy positions or external education.”

Principle 19 of the UNGPs states, “If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it. And if it lacks leverage there may be ways for the enterprise to increase it. Leverage may be increased by, for example, offering capacity-building or other incentives to the related entity, or collaborating with other actors.”

## 34 RECOMMENDATION

**Organize 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.**

This includes different possibilities for content and account tracing, client-side content scanning, and other mitigation techniques for each category of problematic content or accounts, including but not limited to known CSAM, unknown CSAM, content related to grooming and sexual extortion, trafficking, illicit goods sales, misinformation, disinformation, hate speech, and online harassment, etc.

Features such as on-device scanning have already been developed to detect suspicious URLs on WhatsApp.<sup>18</sup> Although each category of problematic content has its own complexities and cannot be easily compared with client-side suspicious link detection, it is still important to explore the feasibility of applying similar techniques for CSAM.

In addition, the Meta Research group should continue to provide grants and calls for research proposals to explore technical and nontechnical mitigation techniques for end-to-end encrypted messaging platforms.<sup>19</sup>

Lastly, the projects carried out by internal and external researchers should take into account a more holistic approach to the human rights impacts of end-to-end encrypted messaging that goes beyond the common privacy against child safety argument.

### EXPLANATION

While some interviewees stated that technical mitigations are possible, others cautioned that proposed mitigations are theoretical and have not been sufficiently tested. This necessitates ongoing discussions with experts in the field as well as additional research on the practical application of such mitigations.

The GNI Principles state, “The development of collaborative strategies involving business, industry associations, civil society organizations, investors and academics will be critical to the achievement of these Principles.”

GNI Principle 5 states that “Participants will take a collaborative approach to problem solving and explore new ways in which the collective learning from multiple stakeholders can be used to advance freedom of expression and privacy.”

GNI Principles Implementation Guideline 4.2 states, “Application Guidance: Promoting rule of law reform could include rule of law training, capacity building with law-related institutions, taking public policy positions or external education.

Principle 19 of the UNGPs states, “If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it. And if it lacks leverage there may be ways for the enterprise to increase it. Leverage may be increased by, for example, offering capacity-building or other incentives to the related entity, or collaborating with other actors.”

## 35 RECOMMENDATION

**Continue to explore ways to responsibly provide data / information for researchers focused on end-to-end encrypted messaging.**

Meta should collaborate with academic and human rights researchers by forming alliances, such as the

<sup>18</sup> <https://faq.whatsapp.com/en/android/26000162/>.

<sup>19</sup> <https://research.fb.com/programs/research-awards/proposals/privacy-preserving-technologies-rfp/>.



Public Interest Research Alliance (PIRA),<sup>20</sup> to support context-specific research projects. In order to build such alliances, Meta should follow privacy protective data sharing agreements, similar to the procedure that the company uses in its Data for Good partnerships.<sup>21</sup>

### EXPLANATION

The GNI Principles state that “The development of collaborative strategies involving business, industry associations, civil society organizations, investors and academics will be critical to the achievement of these Principles.”

GNI Principle 5 states, “Participants will take a collaborative approach to problem solving and explore new ways in which the collective learning from multiple stakeholders can be used to advance freedom of expression and privacy.”

## 36 RECOMMENDATION

**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.**

This research will help Meta understand how users—especially, but not limited to, vulnerable groups including children, women, LGBTQIA+ communities, the elderly, and human rights activists—protect their safety and privacy online and practice resiliency.

For example, one idea proposed during BSR’s external stakeholder interviews was to use “pro-social governance” to address harmful content. Pro-social approaches to social media governance involve encouraging and rewarding behavior, developing community norms for good behavior, and community-based remediation processes rather than punitive actions.<sup>22</sup> This is more applicable to “lower tier” human rights concerns such as the risks of virality of misinformation in messaging platforms.

### EXPLANATION

While much progress has been made on understanding user behavior and tactics on social network platforms, less research has been done on messaging services, especially end-to-end encrypted services. As Meta expands end-to-end encryption, and as its products and features continue to change, there will continue to be a need for ethnographic research to inform product and policy decisions.

The introduction to the UNGPs states, “These Guiding Principles should be implemented in a nondiscriminatory manner, with particular attention to the rights and needs of, as well as the challenges faced by, individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized, and with due regard to the different risks that may be faced by women and men.”

The GNI Principles state, “The development of collaborative strategies involving business, industry associations, civil society organizations, investors and academics will be critical to the achievement of these Principles.”

GNI Principle 5 states, “Participants will take a collaborative approach to problem solving and explore new ways in which the collective learning from multiple stakeholders can be used to advance freedom of expression and privacy.”

<sup>20</sup> See: <https://citrispolicylab.org/pira/>.

<sup>21</sup> See <https://dataforgood.fb.com/approach/> for more information.

<sup>22</sup> <https://law.yale.edu/justice-collaboratory/collab-action-cia/pro-social-media-and-covid-19-disinformation>.

## 37 RECOMMENDATION

**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.**

Meta already funds, pursues partnerships, and collaborates with organizations that protect vulnerable users, including child protection groups. This funding and collaboration is key to addressing many of the human rights risks related to end-to-end encrypted messaging.

At present there are several digital literacy modules and safety tips available on Meta’s Safety Center, but none of these training materials are tailored for protecting users in an end-to-end encrypted environment. By collaborating with civil society organizations, Meta can develop training modules to help users better understand the importance of end-to-end encryption in protecting themselves, in addition to providing guidance about how to keep users safe on any of the end-to-end encrypted platforms. Meta should ensure that these training modules are available in as many languages as possible.

### EXPLANATION

The introduction to the UNGPs states, “These Guiding Principles should be implemented in a nondiscriminatory manner, with particular attention to the rights and needs of, as well as the challenges faced by, individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized, and with due regard to the different risks that may be faced by women and men.”

Principle 19 of the UNGPs states, “If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it. And if it lacks leverage there may be ways for the enterprise to increase it. Leverage may be increased by, for example, offering capacity-building or other incentives to the related entity, or collaborating with other actors.”

The GNI Principles state that “The development of collaborative strategies involving business, industry associations, civil society organizations, investors and academics will be critical to the achievement of these Principles.”

GNI Principle 5 states, “Participants will take a collaborative approach to problem solving and explore new ways in which the collective learning from multiple stakeholders can be used to advance freedom of expression and privacy.”

## 38 RECOMMENDATION

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

The goal of improving child protection online and attacking the internet-wide problem of CSAM is stymied by a lack of ground truth and understanding the true scope of the problem beyond absolute numbers of CSAM. More insight is needed into the amount of unique content, number of victims, timing, trends, among other factors. Understanding and quantifying these factors is key to actually preventing and addressing harm when it occurs, rather than the reactive approach of simply removing and reporting CSAM whenever it is detected.

To do this Meta should work closely with external stakeholders, including researchers, child protection groups, and law enforcement.

## EXPLANATION

The introduction to the UNGPs states, “These Guiding Principles should be implemented in a nondiscriminatory manner, with particular attention to the rights and needs of, as well as the challenges faced by, individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized, and with due regard to the different risks that may be faced by women and men.”

Principle 19 of the UNGPs states, “If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it. And if it lacks leverage there may be ways for the enterprise to increase it. Leverage may be increased by, for example, offering capacity-building or other incentives to the related entity, or collaborating with other actors.”

The GNI Principles state, “The development of collaborative strategies involving business, industry associations, civil society organizations, investors and academics will be critical to the achievement of these Principles.”

GNI Principle 5 states, “Participants will take a collaborative approach to problem solving and explore new ways in which the collective learning from multiple stakeholders can be used to advance freedom of expression and privacy.”

## 39 RECOMMENDATION

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

Meta can engage with organizations and initiatives such as the UNICEF Innovation and Youth and Digital Citizenship+ program at the Harvard Berkman Klein Center to develop tools for children of different ages, caregivers, and educators on topics of safety, reporting mechanisms, communicating and documenting problematic content, and protecting privacy.<sup>23</sup>

The results of such modular training should be publicly accessible through platforms such as Meta’s Safety Center’s Digital Literacy Library, Parents Portal, Youth Portal, and Bullying Prevention Hub. These multimedia materials should be easy to understand in different languages for children of different ages and genders. It is also important that training be tailored to and accessible for differently abled children.

## EXPLANATION

As highly vulnerable rightsholders, children should receive greater protections from the risks of end-to-end encrypted messaging both as users and nonusers. This includes proactive measures by Meta to provide specific training and tools.

Principle 19 of the UNGPs further states, “If the business enterprise has leverage to prevent or mitigate the adverse impact, it should exercise it.”

## 40 RECOMMENDATION

**Create issue-specific working groups within the Safety Advisory Board and among “trusted partners.”**

Such a working group could help address the differing human rights impacts on vulnerable groups in different geographic contexts, and provide Meta with tailored recommendations for emergency assistance.<sup>24</sup> Each working group could focus on issues such as: safety of children aged 0-13; safety of children aged

<sup>23</sup> <https://cyber.harvard.edu/publication/2020/youth-and-digital-citizenship-plus>.

<sup>24</sup> <https://www.facebook.com/help/222332597793306>.

13-18; human trafficking; abuse and online harassment of women; and safety of LGBTQIA+ communities. Appropriate regional representation among working group members would be key to ensuring it is both fair and effective.

### EXPLANATION

The introduction to the UNGPs states, “These Guiding Principles should be implemented in a nondiscriminatory manner, with particular attention to the rights and needs of, as well as the challenges faced by, individuals from groups or populations that may be at heightened risk of becoming vulnerable or marginalized, and with due regard to the different risks that may be faced by women and men.”

The GNI Principles state, “The development of collaborative strategies involving business, industry associations, civil society organizations, investors and academics will be critical to the achievement of these Principles.”

GNI Principle 5 states, “Participants will take a collaborative approach to problem solving and explore new ways in which the collective learning from multiple stakeholders can be used to advance freedom of expression and privacy.”

## 41 RECOMMENDATION

### **Develop innovative methods to categorize reports and summarize their associated metadata for NCMEC.**

Reporting to NCMEC should be more streamlined to minimize the overwhelming flood of reports. Instead of sending out each incident as a stand-alone report, Meta could bundle duplicative reports and provide greater context that informs NCMEC and law enforcement’s ability to prioritize cases. These types of effective and precise bundling methods will help NCMEC and law enforcement to do their job more efficiently and effectively.

### EXPLANATION

Multiple interviewees expressed that law enforcement agencies are often overwhelmed by the number of reports they receive from NCMEC. Each report is generated individually, so thousands of reports may reference the same piece of content. This “noise” makes it hard for NCMEC and law enforcement to filter out quality reports and prioritize cases to pursue. If Meta can help minimize the “noise,” it will improve NCMEC and law enforcement’s ability to take action on priority reports.

The GNI Principles state, “The development of collaborative strategies involving business, industry associations, civil society organizations, investors and academics will be critical to the achievement of these Principles.”

GNI Principle 5 states, “Participants will take a collaborative approach to problem solving and explore new ways in which the collective learning from multiple stakeholders can be used to advance freedom of expression and privacy.”

GNI Principles Implementation Guideline 4.2 states, “Application Guidance: Promoting rule of law reform could include rule of law training, capacity building with law-related institutions, taking public policy positions or external education.

## 42 RECOMMENDATION

### **Continue to actively work with trafficking organizations that have built relationships with survivor communities.**

Trafficking is a very nuanced and industry-focused issue that spans various industries such as tourism and traveling, beauty and health services, fashion, and sex work.

To act effectively in creating any tools or resources to detect malicious activities and protect potential victims, it is important for Meta to continue supporting organizations that already have capacity and understanding about context and together help develop detection tools and mitigation techniques.

### EXPLANATION

As highly vulnerable rightsholders, trafficking victims should receive greater protections from the risks of end-to-end encrypted messaging. This includes proactive measures by Meta to build and provide tools and techniques that mitigate and remedy such human rights impacts.

The GNI Principles state, “The development of collaborative strategies involving business, industry associations, civil society organizations, investors and academics will be critical to the achievement of these Principles.”

GNI Principle 5 states, “Participants will take a collaborative approach to problem solving and explore new ways in which the collective learning from multiple stakeholders can be used to advance freedom of expression and privacy.”

## 43 RECOMMENDATION

### **Proactively collaborate with, train, and inform law enforcement about how to achieve their objectives in end-to-end encrypted contexts in a rights-respecting way.**

Showing law enforcement and relevant policymakers what can be achieved without access to message content can help effectively push back against regulatory requirements that weaken or break encryption.

Governments may begin to request more user data in end-to-end encrypted contexts. Meta can mitigate government overreach in data requests through engagement and training of law enforcement about what data is relevant and how it can be used in criminal investigations.

This includes continuing working with civil society organizations such as ICMEC that have the knowledge and access to law enforcement groups in different countries to help provide training, especially for issues around child exploitation and human trafficking.

Collaboration need not be seen as obligatory and “all or nothing.” Rather, it should be done on a case-by-case basis, based on the rule of law context of the jurisdiction involved, and have limited objectives and be appropriately scoped to prevent misuse of new capabilities or related adverse human rights impacts.

### EXPLANATION

This recommendation is informed by the counterbalancing exercise conducted earlier in this report on the tension between the right to privacy and the right to security. This exercise suggested that in order to soften the impacts of end-to-end encrypted messaging on the right to security, Meta should proactively collaborate with law enforcement on the detection and prosecution of digital crimes by educating them about the types of digital evidence available to them and providing them with the tools to make sense of it.

The GNI Principles state, “The development of collaborative strategies involving business, industry associations, civil society organizations, investors and academics will be critical to the achievement of these Principles.”

GNI Principle 5 states that “Participants will take a collaborative approach to problem solving and explore new ways in which the collective learning from multiple stakeholders can be used to advance freedom of expression and privacy.”

GNI Principles Implementation Guideline 4.2 states, “Application Guidance: Promoting rule of law reform could include rule of law training, capacity building with law-related institutions, taking public policy positions or external education.

## 44 RECOMMENDATION

### **Meta should continue working with other social media and internet companies to explore techniques to mitigate potential human rights impacts of end-to-end encrypted messaging.**

Both PhotoDNA (in the case of the sexual abuse and exploitation of children) and GIFCT’s Hash Sharing Consortium and URL-sharing (in the case of terrorism and digital recruiting) are the results of cross-company collaborations that may be impacted by Meta decisions regarding end-to-end encrypted messaging.

Even if Meta is unable to contribute to such databases as significantly post-expansion of end-to-end encryption, it should continue sharing knowledge and best practices.

Meta should also work with other companies in building industry standards based on human rights principles, especially in the fight against grooming of children and human trafficking in situations where national laws are inconsistent or inadequate.

It is important for such industry collaborations to be transparent and to be accompanied by independent oversight and accountability mechanisms.

Lastly, Meta should consider engaging with companies outside the tech industry to counter potential misuse of end-to-end encrypted messaging services. For example, Meta could work with banks and payment providers to identify and track bad actors using Meta services for illicit sales.”

### EXPLANATION

The GNI Principles state, “The development of collaborative strategies involving business, industry associations, civil society organizations, investors and academics will be critical to the achievement of these Principles.

GNI Principle 5 states, “Participants will take a collaborative approach to problem solving and explore new ways in which the collective learning from multiple stakeholders can be used to advance freedom of expression and privacy.”

## Public Communications

## 45 RECOMMENDATION

### **Meta should publicly communicate its strategy and action plan to mitigate the adverse human rights impacts of end-to-end encrypted messaging, including progress against these recommendations over time.**

Part of addressing the potential adverse human rights impacts of expanding end-to-end encryption is

publicly “knowing and showing” what Meta is doing to mitigate those impacts. This communication can take place in a number of places, from publication of an annual human rights report to publicly available product policies. It should also take place on an ongoing basis as Meta assesses the effectiveness of its mitigations and evolves approaches over time.

#### EXPLANATION

Principle 21 of the UNGPs states that in order to account for how they address their human rights impacts, business enterprises should be prepared to communicate this externally, particularly when concerns are raised by or on behalf of affected stakeholders.

Business enterprises whose operations or operating contexts pose risks of severe human rights impacts should report formally on how they address them.

In all instances, communications should: (a) Be of a form and frequency that reflect an enterprise’s human rights impacts and that are accessible to its intended audiences; (b) Provide information that is sufficient to evaluate the adequacy of an enterprise’s response to the particular human rights impact involved; (c) In turn, not pose risks to affected stakeholders, personnel, or legitimate requirements of commercial confidentiality.



## About BSR

BSR™ is an organization of sustainable business experts that works with its global network of the world's leading companies to build a just and sustainable world. With offices in Asia, Europe, and North America, BSR™ provides insight, advice, and collaborative initiatives to help you see a changing world more clearly, create long-term business value, and scale impact.

[www.bsr.org](http://www.bsr.org)