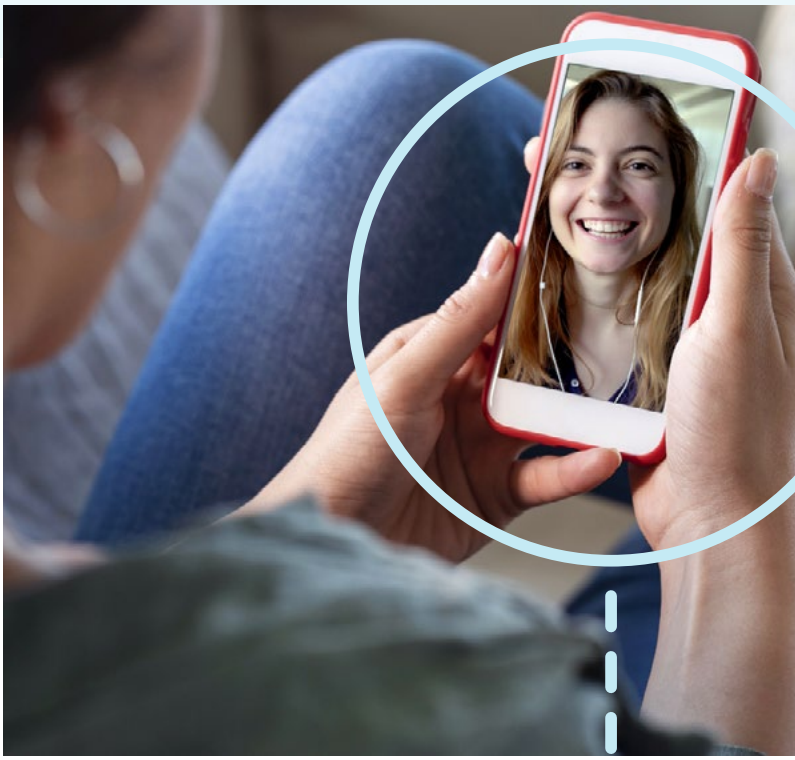
BSR®

HUMAN RIGHTS IMPACT ASSESSMENT

Meta's Expansion of End-to-End Encryption

EXECUTIVE SUMMARY

Contents



KEY POINTS

End-to-end encryption is increasingly important to protecting 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 the context of rising digital authoritarianism, increasingly sophisticated digital security threats that place both individuals and key industries and infrastructure at risk, and the growth of sensitive communications online and across geographic borders.

Privacy and security while using online platforms should not only be a privilege of the technically savvy and those able to make proactive choices to opt into end-to-end encrypted services, it should be something that is democratized and available to everyone.

A human rights approach can bring needed nuance to the broader encryption policy debate.

Today's encryption debate pits two opposing groups against each other, with privacy on one side and security on the other. However, the reality is much more nuanced, with privacy and security concerns on both sides, and many other human rights that are impacted both positively and negatively. The holistic view taken in this assessment reveals this nuance by considering the potential impacts of end-to-end encryption on all human rights, as well as the connectivity between rights.

Expanding end-to-end encryption across Meta's messaging platforms will address adverse human rights impacts arising from the absence of ubiquitous end-to-end encryption today.

End-to-end encryption of messaging directly enables the right to privacy, which in turn enables other rights such as freedom of expression, association, opinion, religion, movement, and bodily security. Meta's expansion of end-to-end encrypted messaging will therefore result in increased realization of these rights.

Meta would be "directly linked"[1] to potential adverse human rights impacts associated with the expansion of end-to-end encryption.

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. This does not mean that Meta is not responsible for addressing these harms; rather, understanding Meta's relationship to harm provides insight into the leverage Meta has to address it.

BSR's assessment is that in and of itself, end-to-end encryption does not "cause" or "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.

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 causing or "contributing" to) the potential adverse human rights impacts associated with the expansion of end-to-end encryption.

The expansion of end-to-end encryption involves challenging human rights trade-offs with wider system-level implications.

One of the most challenging debates related to end-to-end encryption is whether companies should use nascent "client-side scanning" techniques to scan messages to detect and report child sexual abuse material (CSAM). However, because those same techniques can be used to detect a wide variety of content, that question is part of a larger debate about content moderation in private messaging in general, as well as in and end-to-end encrypted services.

There is currently no consensus on where to draw the line on content moderation in a messaging context. With the existence of large group messages, messaging platforms can sometimes seem like a quasi-public space and face many of the same content issues seen in open social media platforms. However, messaging is still largely a private space, and moderation of anything other than content that always and clearly constitutes a human rights violation (such as CSAM) would be an unnecessary and disproportionate infringement on privacy and freedom of expression.

The technical feasibility, resiliency, and integrity of client-side scanning methods for end-to-end encrypted messaging at scale is uncertain and highly debated. Even if feasible, implementing client-side scanning to detect CSAM risks an irreversible slippery slope. Government regulation of online content is increasing around the world, both in the legitimate pursuit of safe and rights-respecting online spaces and in the illegitimate pursuit of censorship and oppression. There is a significant risk that a well-intentioned attempt to protect children would be abused by governments to require Meta to block and report legitimate content that a government dislikes. This would lead to the unjust restriction 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.

There are no easy answers to addressing the trade-offs. 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 conflicts, but the fast moving nature of the slippery slope risk makes that challenging.

More due diligence is needed on potential future mitigation measures before decisions with lasting consequences are made.

We conclude that Meta should 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. We note that the only client-side scanning method proposed thus far that may potentially meet these requirements is homomorphic encryption, which allows for the processing of data in its encrypted state. However, it is not yet technically feasible to implement in messaging at scale, and therefore our analysis and conclusions about homomorphic encryption are speculative.

If Meta identifies technically feasible client-side scanning methods capable of detecting CSAM while maintaining the cryptographic integrity of end-to-end encryption, then it 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 could be adequately mitigated.

1

Project Context and 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."[2]

In this post Zuckerberg described the challenges of balancing privacy and safety in the context of end-to-end encryption, and stated that Meta will continue to discuss these challenges with experts before fully implementing end-to-end encryption across Meta's messaging platforms.

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 Instagram DMs does not offer end-to-end encrypted messaging capabilities (though at the time of writing optional end-to-end encrypted messaging is being publicly tested). 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) 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 Meta's messaging services, using a methodology

based on the UN Guiding Principles on Business and Human Rights (UNGPs). The objectives of this assessment 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 and standards;
- **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.

It is important to note that this assessment has been undertaken in parallel with Meta's decision-making about how to expand end-to-end encryption to all messaging services and make them capable of cross-app communication. This deliberate integration of human rights into the design and decision-making phase of product and feature development 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 assessment does not include "final state" review of human rights and end-to-encryption in Meta's messaging services.[3]

It should also be noted that the full BSR assessment contains a far more detailed, nuanced, and thorough analysis of this very complex topic. This executive summary necessarily focuses on the key points only.

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 establishing 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 discussed.

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 conducting this assessment, BSR relied 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.

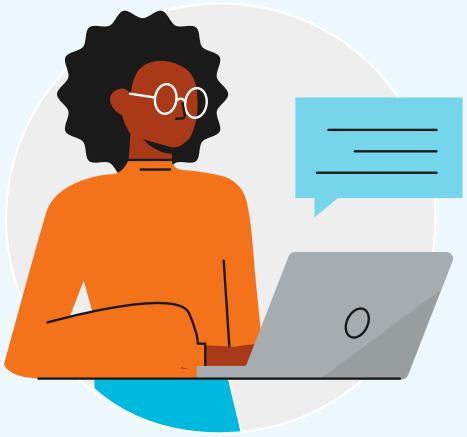
2

What Is End-To-End Encryption?

End-to-end encryption scrambles messages in such a way that only the sender and the recipient can decipher them. Messages are encrypted on the device of the sender and decrypted on the device of the recipient, and even Meta, the company providing the messaging service, cannot view the contents of messages.[4]

It is difficult for third parties to gain access to the content of communications made using end-to-end encryption.[5] 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 have hacked into the device itself via spyware or other means.[6] For this reason end-to-end encrypted messaging is considered the most secure and privacy-protective method of communication.[7]

How End-to-End Encrypted Messaging Works



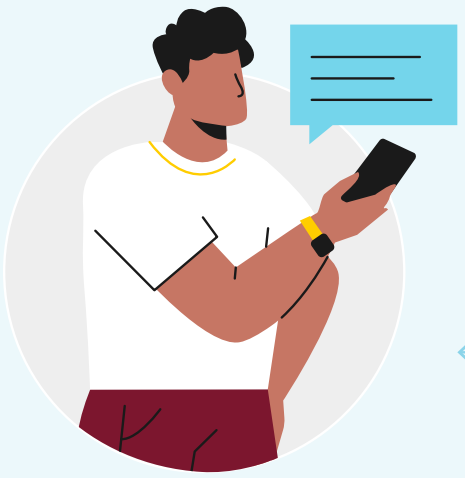
Sender

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.

3

Why Is End-To-End Encryption Relevant For 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 role in society's strategies to protect, respect, and fulfil human rights in today's political, social, and technical context.

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. 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.

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.

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

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.[8]

The strategies and tactics of authoritarianism are increasingly taking place online through surveillance, spyware, and other methods to turn online spaces into more hostile environments. Freedom on the Net 2021 found that global internet freedom declined for the 11th consecutive year, and more governments than ever before arrested users for nonviolent political, social, or religious speech. Freedom on the Net 2021 also found that authorities in at least 45 countries were suspected of obtaining sophisticated spyware or data-extraction technology from private vendors, while Freedom on the Net 2019 found 40 of the 65 countries studied had instituted advanced social media monitoring programs.[9]

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 staying in touch with friends and families spread around the world, more of our private

communications than ever before are taking place over platforms, apps, and services that rely on encryption to keep them secure.

Our communications and networks 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 by governments that are not.

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 a privilege of the technically savvy and those able to make proactive choices to opt into end-to-end encrypted services, it should be something that is democratized and available to everyone.

4

How Does A Human Rights Approach Help 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.

This debate sets two opposing groups against each other in the name of two potentially competing human rights—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, traffickers, and child abusers, and makes it more difficult 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 is therefore important that Meta and other relevant actors address them in an informed, deliberate, and thoughtful manner.

The purpose of this assessment is to take this holistic view by considering the potential impacts of end-to-end encrypted messaging on a wide range of human rights, and how adverse impacts should be addressed by Meta and other actors. The assessment therefore incorporates four main elements:

We considered impacts on all human rights. In addition to privacy and security, this assessment considers the impact the expansion of end-to-end encryption will have on the universe of rights

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 is therefore important that Meta and other relevant actors address them in an informed, deliberate, and thoughtful manner.

codified in international human rights instruments, including 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), as well as many other relevant international human rights instruments.

We assessed the connectivity between rights. Human rights impacts are interconnected and interrelated. However, human rights can be in tension with one another for legitimate reasons, and rights-based methods can be used to define a path forward when two conflicting 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"[10] 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 emphasized the interests of vulnerable groups. We paid special consideration to identifying and addressing the specific needs of vulnerable groups who face heightened risks, or different risks, 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 negative human rights impacts of end-to-end encrypted messaging, but they may also stand to gain the most from the human rights benefits. To better understand the impact of end-to-end encrypted messaging on vulnerable groups we engaged with independent stakeholders and experts, such as academics and civil society organizations specializing in privacy, freedom of expression, protection of human rights defenders, child rights, counterterrorism, human trafficking, and violence against women.

We considered the roles and responsibilities of different actors. Human rights opportunities that arise from deploying end-to-end encryption across Meta's messaging services occur when Meta's platforms are used as intended. In contrast, the human rights risks arising from end-to-end encryption tend to be associated with the misuse or abuse of Meta's platforms by bad actors who disregard terms of service, violate the law, and adversely impact the rights of others. This has significant implications for which actors hold what responsibility to address potential adverse human rights impacts, and emphasizes the need to take system-wide approaches that consider technology in its broader societal context.

5

What Are The Key Human Rights 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.

Encryption Context

Meta's expansion of end-to-end encryption 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. End-to-end encryption of messaging 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

End-to-end encryption of messaging 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.

service terms that they have agreed to. However, this does not mean that Meta is not responsible for addressing these harms; rather, Meta's relationship to harm provides insight into the leverage Meta has to address them.

As the parent company of 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 harm at 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, sexual abuse, and exploitation of children; exchanging illegal goods and content; or sharing 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 messaging platforms. Sophisticated technology use is increasingly part of criminal tradecraft, and the percentage of criminals with the knowledge and skills to use end-to-end encryption will continue to increase over time. For this reason, choosing not to provide end-to-end encryption would likely not result in an improved 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.

User expectations, and therefore informed consent, varies based on the messaging platform used. User expectations differ for Messenger and Instagram, which started as open social network platforms, compared to WhatsApp, which has always been a private messaging app. This makes a notable difference when it comes to informed consent across a range of topics (such as privacy and content policies), as well 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 better human rights outcomes over time.

Content removal is just one way of addressing harms. Prevention methods are feasible in an end-to-end encrypted environment and are essential for 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. Content removal is important for 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. Meta can proactively prevent harm 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.

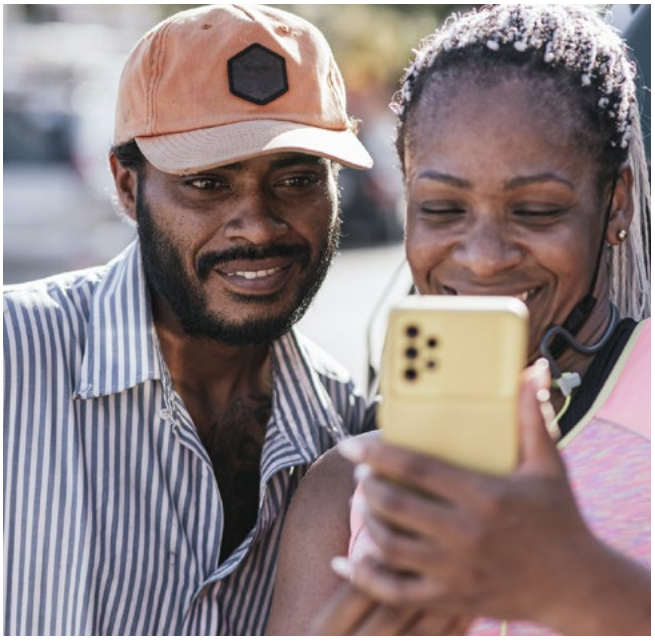
Meta can proactively prevent harm in end-to-end encrypted messaging through the use of behavioral signals, public platform information, user reports, and metadata to identify and interrupt problematic behavior before it occurs.

There is no consensus on the degree of content moderation companies should undertake on messaging services. While there is increasing consensus about content moderation boundaries for content posted to open platforms (such as Meta and Instagram), this has not yet extended to the messaging context (including WhatsApp, but also SMS services, other stand-alone messaging apps, and live audio calls and video calls). This dilemma will be especially relevant for Meta given the different content policies that currently apply across the three messaging platforms.

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 factors include languages, local information ecosystems, and type of devices available. As a result, the risks and opportunities of end-to-end encrypted messaging are likely to be amplified in some locations.

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. By contrast, in countries without extensive surveillance regimes but with significant ethnic or communal 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 in others, and certain types of problematic content, such as child sexual abuse material, 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 are disproportionately impacted by the actions of others, such as

authoritarian governments or other bad actors with nefarious intent. A human rights-based approach to end-to-end encrypted messaging therefore needs to pay special attention to the circumstances of vulnerable groups—such as those who use lower quality devices (e.g., devices with less power/processing capability); have lower levels of digital literacy; and use languages that Meta does not support. It should also reflect the reality that significant numbers of children under the age of 13 use private messaging services, despite minimum age requirements. Approaches to end-to-end encryption need to be designed with a wide range of users in mind, not simply those over the age of 13 in affluent markets or circumstances.

Approaches to end-to-end encryption need to be designed with a wide range of users in mind, not simply those over the age of 13 in affluent markets or circumstances.

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 some regions may have more in-country personnel, language and translation services, moderation capacity, or technical interventions than others.

Product Policy

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.[11] 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—that could allow the detection of harmful content such as CSAM.

Because homomorphic encryption could maintain 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.[12]

Meta will not be able to proactively review messages for content that violates its content policy standards, so 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.

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 a 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.

There are important choices to be made about what content policies apply in an end-to-end encrypted messaging environment. Meta'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 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; second, whether, in the context of cross-app communication, content standards should be consistent across the three messaging platforms.

Product Factors

In an end-to-end encrypted messaging environment user reporting of problematic content and accounts is a critically important enforcement mechanism. Meta will not be able to proactively review messages for content that violates its content policy standards, so 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. Given the challenges of scale, speed, and volume, it will be impossible for a "perfect" reporting channel and appeals mechanism 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) provide 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

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.

prevent human rights harm by discouraging users from sharing problematic content or engaging in abusive behavior.

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 both preventing and identifying misuse, high-risk behavior, and threat actors in an end-to-end encrypted environment. However, mass collection and analysis of metadata also presents significant privacy risks, which need to be carefully weighed and addressed. 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.

Using machine learning (ML) systems to detect and prevent problematic content is important for harm prevention and response at the scale of Meta, but on its own is not sufficient. 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. Although Meta should invest in improving the quality of its ML classifiers, adequate human review resources across geographic and linguistic contexts also need to 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,[13,14] 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.[15] 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 significant, disproportionate restrictions on a range of rights, and should therefore not be pursued.

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. Nevertheless, research into these 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 that would need to be addressed. For example, if Meta starts detecting and reporting universally illegal content like CSAM, governments may 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.[16] 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. It is also 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 and other rights, then client-side scanning should not be pursued.

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

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.[17]

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 human rights impacts associated with cross-app communication 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.

External Engagement

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, in today's world law enforcement also benefits from vastly more data and advanced data analysis capabilities than in the past and may not be faced with a net loss in capability overall. 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. A case can be made that Meta should proactively support 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—as part of its responsibility to address human rights harm connected with end-to-end encrypted messaging. However, in a growing number of cases, government intentions are not aligned with human rights or there is a lack of rule of law, 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 that access to actual message content may provide.

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-versus-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 engage productively with other government officials to inform them of Meta's approach to assisting law enforcement and all the ways in which evidence and intelligence gathering can successfully 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.

6

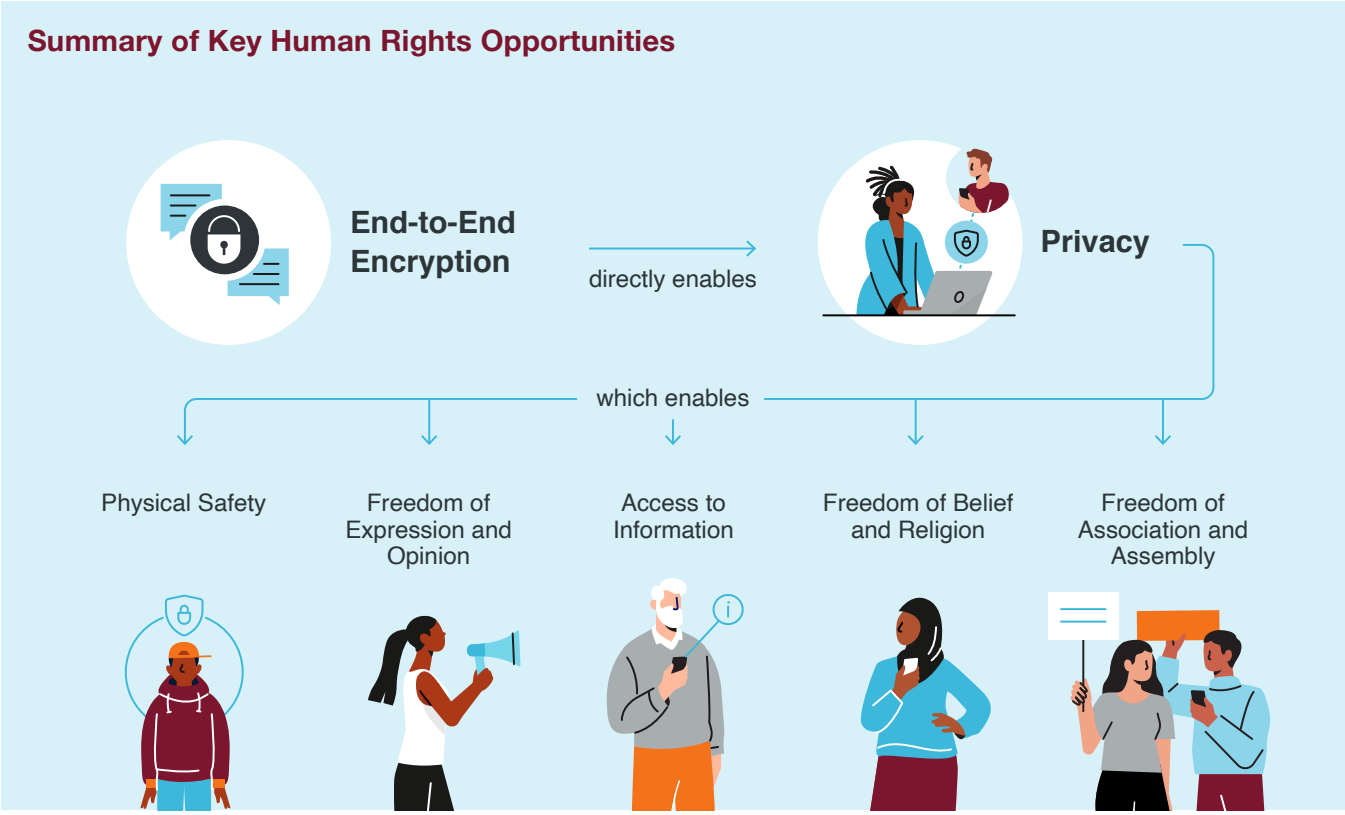
What Are The Main Human Rights Opportunities And Risks?

The full BSR assessment provides a thorough list of all the human rights potentially impacted by Meta's expansion of end-to-end encryption. In this summary we highlight the most significant opportunities and risks.

According to the UNGPs, Meta cannot "offset" the opportunities against the risks or conclude that there is a "net benefit" to human rights arising from the expansion of end-to-end encryption, and therefore ignore the harms. Meta's responsibility according to the UNGPs is to address all the adverse human rights impacts with which it is involved, regardless of the benefits.

It is 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, and because users can use end-to-end encrypted messaging in ways that harm nonusers.

Our recommendations in Section 7 are intended to describe appropriate action for Meta to address the human rights risks and maximize the opportunities.



Human Rights Opportunities

The expansion of end-to-end encryption will directly result in the increased realization of a range of human rights—or, to reframe, the expansion of end-to-end encryption will address many human rights risks associated with the absence of ubiquitous end-to-end encryption on messaging platforms today.

Privacy and its knock-on benefits: End-to-end encryption of all messaging platforms will increase the privacy of all users, not just those who know or care about encryption enough to opt-in to end-to-end encrypted platforms. This in turn will enable and reinforce other human rights—when the right to privacy is respected, people can more freely exercise other rights that depend on privacy, such as expression, opinion, association, movement, religion, and belief, among many others.

Freedom of expression and opinion, belief and religion, and association and assembly, and access to information: By ensuring the privacy of communications, end-to-end encrypted messaging enables people to freely form opinions, express themselves, share information, associate, and assemble without fear of retribution.

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. 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 adversarial governments or citizens. This can

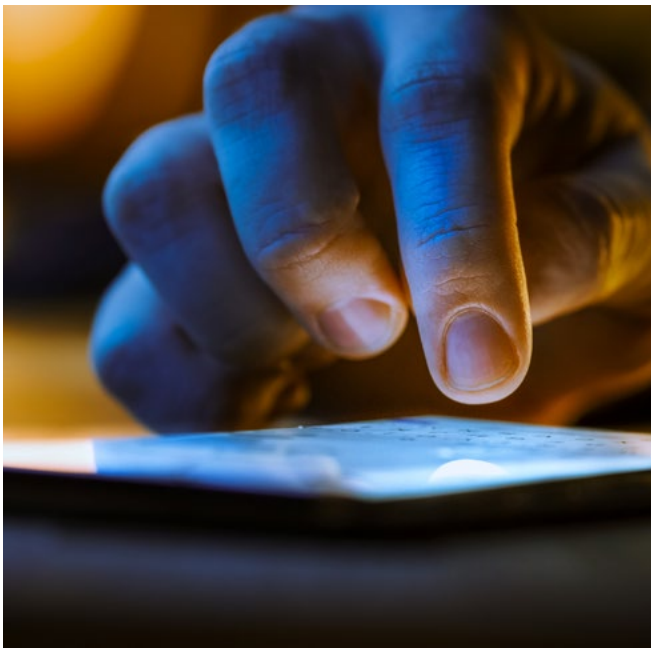
benefit other rights such as freedom from torture, degrading treatment, or punishment, and freedom from exploitation, violence, and abuse.

Child rights: Although there are significant child rights risks arising from the expansion of end-to-end encryption (see 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.

Access to remedy: 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 remedy.

Culture and science: The privacy protections of end-to-end encrypted messaging can enable community members to maintain cultural ties in contexts where their culture is socially or legally repressed. By being present in widely used messaging platforms, end-to-end encryption across all three of Meta's messaging platforms would enable more people to enjoy the benefits of scientific advancement.

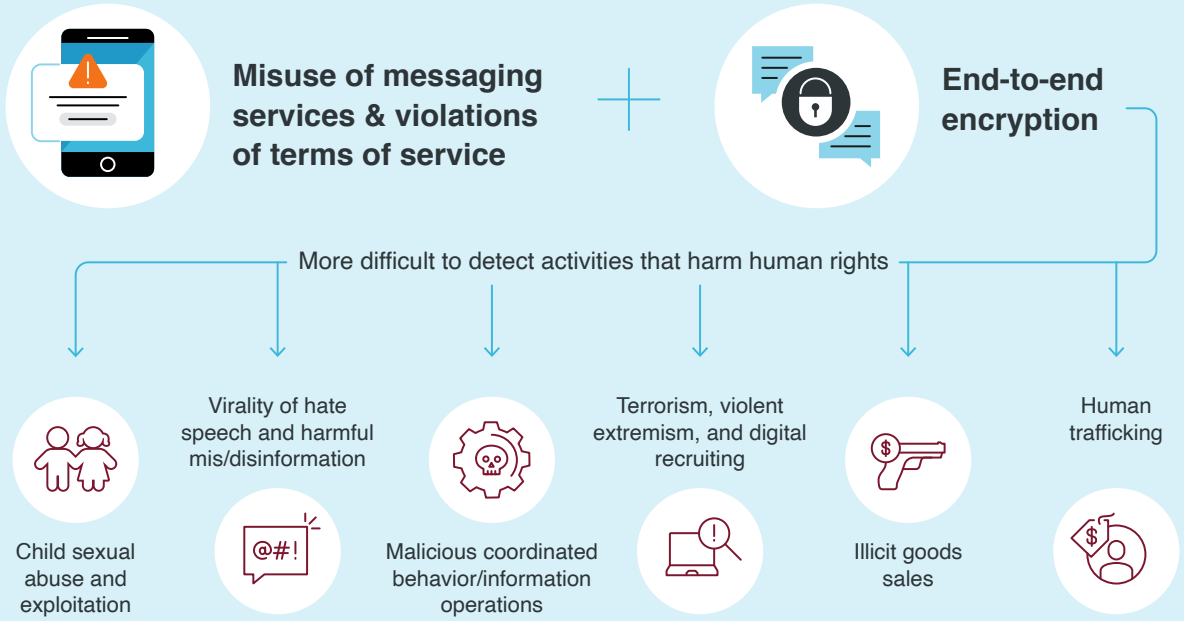
Work, equal pay, and fair wages: The privacy protections of end-to-end encrypted messaging and its benefits for free association can enable and protect labor union communication, recruitment, and activity in places and contexts where labor rights are restricted.



Participation in government: The privacy protections of end-to-end encrypted messaging can enable people to more freely and safely discuss and facilitate participation in government in situations where there are attempts to interfere with free and fair elections.

Arbitrary arrest and exile: The privacy protections of Meta's end-to-end encrypted messaging platforms may help 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.

Summary of Key Human Rights Risks



(A range of appropriate mitigations are suggested to address these risks)

Human Rights Risks

In contrast to the opportunities, 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. This is because end-to-end encryption makes the human rights risks of messaging platforms more difficult to detect.

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 relatively targeted.

Child Sexual Abuse and Exploitation: The expansion 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.

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 challenging to detect in an end-to-end encrypted environment and therefore may make it more difficult to address potential harm.

Malicious Coordinated Behavior/Information Operations: Malicious coordinated behavior, both authentic (i.e., by real people using real

accounts) and inauthentic (i.e., by people using fake accounts), can undermine the integrity of social media platforms and messaging services. While coordinated behavior is not in itself a human rights violation (and can also be carried out legitimately as part of human rights campaigns, for example), it can enable bad actors to exploit 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.

Illicit Goods Sales: In addition to legal commerce, a wide range of illicit activity takes place via encrypted messaging services, such as activities related to weapons, drugs, or cyber-fraud services. These activities can impact bodily security rights, and may be more difficult to detect and address in an end-to-end encrypted environment.

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.

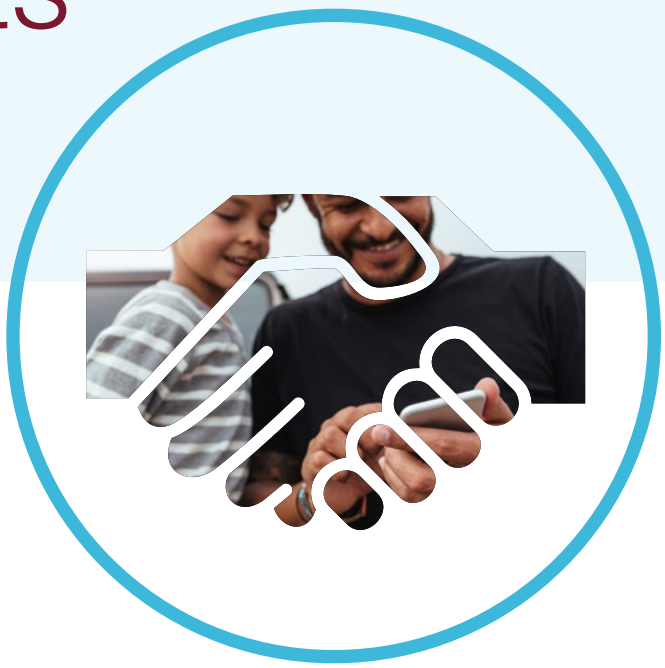
Terrorism, Violent Extremism, and Hate Groups: Violent extremist and terrorist groups have proven to be tech savvy and have increasingly used end-to-end encrypted messaging platforms to communicate with followers, disseminate propaganda, incite violence, and coordinate terrorist attacks that result in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

Nondiscrimination: Content that intends to harass users, based on characteristics such as gender, religion, ethnicity, LGBTQIA+ status, or political views, may be shared on end-to-end encrypted messaging platforms, but not reported or removed. In addition, the use of behavioral signals and metadata analysis (in the absence of access to content) may result in law enforcement actions that are discriminatory in nature—for example, during counterterrorism efforts.

Privacy: End-to-end encrypted messaging may be used to share content that violates people's privacy, such as nonconsensual intimate images. Making all messaging platforms capable of cross-app communication may enable people to find users on different platforms, increasing "discoverability" and risking the privacy of users who do not have accounts on all platforms 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, such as human rights defenders, whistleblowers, and journalists.

7

What Are The Main Human Rights Trade-Offs?



The full BSR assessment provides a thorough analysis of how potentially competing human rights—such as privacy and child safety—can be addressed in ways that allow the fullest expression of both rights. Here we highlight a few prevalent themes.

The detection of child sexual abuse material has larger implications. One of the most challenging debates related to end-to-end encryption is whether companies should scan messages to detect and report child sexual abuse material. Currently, Meta scans messages to detect and report known CSAM on its unencrypted messaging platforms as part of an industry-wide effort in collaboration with government authorities and civil society. It also scans unencrypted content in WhatsApp such as profile photos, group names and descriptions, and user reports, which it will still be able to do with the expansion of end-to-end encryption. While scanning and removal of known CSAM is not a catch-all solution for preventing child sexual abuse online, it is a mitigation against the revictimization of pictured victims and assists in the identification of those distributing CSAM. To continue to scan message content in an end-to-end encrypted messaging context, Meta would need to use one of several nascent hash-based solutions often collectively referred to as "client-side scanning."

The debate about whether companies should scan messages to detect CSAM in end-to-end encrypted messaging is part of a larger debate about content moderation in private messaging and end-to-end encrypted services

The debate about whether companies should scan messages to detect CSAM in end-to-end encrypted messaging is part of a larger debate about content moderation in private messaging and end-to-end encrypted services. Hash-based approaches used to detect CSAM could also be used to detect, block, or remove many other kinds of problematic content, such as nonconsensual intimate images, hate speech, and terrorist content. This means that the debate about CSAM detection raises a number of other challenging dilemmas for which there is no easy answer.

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.

There is no consensus on where to draw the line on content moderation in a messaging context. Increasingly, messaging platforms are no longer just private domains. The existence of large WhatsApp groups[18] (although they are a very small percentage of overall chats) means that private messaging can sometimes feel 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 without algorithmic promotion of content.

The impacts of viral misinformation, hate speech, and other types of problematic content on WhatsApp has led some to call for the same content moderation practices implemented on social media platforms to be applied to messaging platforms as well. However, despite these trends, messaging platforms are still largely private spaces, and moderation of anything other than the most egregious types of content would be an unnecessary and disproportionate infringement on privacy and freedom of expression. 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 service being provided.

Hash-based systems that operate in an end-to-end encrypted environment would be a 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. 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 the removal of too much legitimate content, constituting an undue burden on freedom of expression.

The limitations of hash-based content moderation and the risks of over-enforcement leads BSR 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[19] because both constitute live violations of people's privacy when shared, and both can be included in a hash database. While many other types of content—such as hate speech and 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.

The technical feasibility, resiliency, and integrity of client-side scanning methods is uncertain. 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 far too computationally intensive to implement on a large-scale messaging platform today, even for high-end mobile devices. For example, Meta's own research of a homomorphic encryption approach—the only approach developed thus far that could maintain the cryptographic integrity of end-to-end encryption—found that it would take around 20 million seconds (equivalent to over seven months) to run on each message.

Therefore, even if cryptographic integrity-maintaining client-side scanning for CSAM in end-to-end encrypted messaging were technically feasible, 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.

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. There is the risk that bad actors may take advantage of the technical vulnerabilities of these solutions to game the system by, for example, creating false negatives to enable violating content to pass, creating false positives to erroneously flag non-violating content, or using unofficial clients to deactivate the code running client-side scanning.

The "slippery slope" risk of CSAM detection. Government regulation of online content has grown enormously in recent years, both in the legitimate pursuit of safe and rights-respecting online spaces and the illegitimate pursuit of censorship and oppression.[20] Therefore, even if cryptographic integrity-maintaining client-side scanning for CSAM in end-to-end encrypted messaging were

technically feasible, 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 would undoubtedly lead to the unjust restriction 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. Even some who had previously proposed client scanning approaches have subsequently altered their view for these reasons.[21]

In this context, 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.

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, it is BSR's view that client-side scanning should not be pursued.

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 preserve 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, it is BSR's view that 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 would constitute a disproportionate restriction on the freedom of expression rights of all users.

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.

8

What Are BSR's Recommendations For Meta?

According to the UNGPs, Meta has a responsibility to (a) avoid "causing" or "contributing" to adverse human rights impacts through its own activities, and address such impacts when they occur; and (b) seek to prevent or mitigate adverse human rights impacts that are "directly linked" to its products or services by its business relationships, even if it has not "contributed to" those impacts.

It is BSR's view that **expanding end-to-end encryption across Meta's messaging platforms will address adverse human rights impacts arising from today's absence of ubiquitous end-to-end encryption and result in the increased realization of a diverse range of human rights**, including privacy, physical safety, freedom of expression, freedom of assembly and association, access to remedy, and participation in government. Expanding end-to-end encryption is an important step in addressing adverse human rights impacts with which Meta is already involved.

However, a key question arising in this assessment is whether the expansion of end-to-end encryption will also "cause, contribute to, or be directly linked to" new adverse human rights impacts, and how these potential adverse human rights impacts should be addressed.

Expanding end-to-end encryption across Meta's messaging platforms will address adverse human rights impacts arising from today's absence of ubiquitous end-to-end encryption and result in the increased realization of a diverse range of human rights

BSR's assessment is that in and of itself, end-to-end encryption does not "cause" or "contribute to" (i.e., enable, facilitate, incentivize, or motivate) harm.

BSR's assessment is that ***in and of itself, end-to-end encryption does not "cause" or "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. For this reason, ***BSR concludes that Meta would be "directly linked" to the majority of potential adverse human rights impacts*** associated with the expansion of end-to-end encryption.

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 considered "contributing" to the increased harm, but only if reasonable mitigation measures are not put in place. ***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.***

However, BSR would like to note that this assessment, and the particular mix of risks and opportunities it has surfaced, has been very demanding, fraught with ethical challenges, complex dilemmas, and difficult trade-offs for which there are no easy answers and for which human rights arguments can be made on either side. Ultimately, BSR's conclusions and recommendations reflect the reality of an industry-wide shift toward end-to-end encryption, and the need to take appropriate actions to address the risks associated with this shift.

It is also 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 positioned 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.

In this context, BSR makes the following 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. By implementing BSR's recommendations, Meta will also contribute to the remedy ecosystem for many of the harms associated with end-to-end encrypted messaging—for example, by supporting entities that help victims of harm access justice and remediation services. Meta will also prepare itself to provide remedy to victims as well, such as through improved user reporting channels.

BSR's recommendations are divided 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. BSR's full report includes a more detailed description of each recommendation accompanied by a human rights-based rationale.

Recommendations

Product

Recommendations about specific products and features, such as reporting, account linking, and discoverability.

RECOMMENDATIONS

- Provide more consistent, cohesive, and accessible methods for user reporting across messaging platforms.
- Ensure that user interfaces (especially user reporting features) are easy to find, simple to use, and available in all the languages Meta supports.
- To protect children from unsolicited interactions with adults (which might lead to grooming and trafficking), Meta's UX/UI Research group should conduct participatory and co-design workshops to test user reporting features with children.
- Develop documentation and measurement techniques to assess the degree to which user reporting is helping to keep users safe online.
- Explore and define how to verify the authenticity of user reports.
- Invest in processes to ensure that users who have violated platform policies cannot return.
- Expand and simplify in-app support and education features for vulnerable groups, such as children or those with lower levels of digital literacy.
- Assess options for "friction" when contacting groups and strangers on messaging platforms in order to minimize unsolicited interactions, virality of harmful mis/disinformation and hate speech, harmful coordinated behavior, and other actions that may lead to adverse human rights impacts.
- 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.
- 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.

Process

Recommendations for how Meta can detect and mitigate human rights risks, such as user reporting and behavioral signals.

RECOMMENDATIONS

- Continue to invest in harm prevention strategies in end-to-end encrypted messaging, such as the use of metadata analysis and behavioral signals, redirection/behavioral nudges, user education, etc., and communicate publicly on lessons learned and the effectiveness of such methods for addressing different kinds of harm.
- During the design and development of ML techniques to proactively detect harmful accounts and content in end-to-end encrypted messaging, follow "human rights by design" guidelines to ensure user privacy, fairness, transparency, interpretability, and auditability.
- 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, such as accessible user reporting features; user education; metadata analysis; use of behavioral signals; further investigation of scalable and cryptographic integrity-respecting client-side scanning techniques for CSAM; law enforcement training and partnerships; civil society partnerships; and the development of metrics to quantify the scope of CSAM and corresponding harm, among others.
- 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.
- If Meta identifies client-side scanning methods capable of detecting CSAM while maintaining the cryptographic integrity of end-to-end encryption, then this should only be implemented after a review of the potential adverse human rights impacts (for example, on privacy, freedom of expression) and a conclusion that those impacts could be adequately addressed
- Conduct human rights due diligence on cross-app communication.

Product Policy

Recommendations for product policy across products, such as Community Standards.

RECOMMENDATIONS

- Develop new privacy policies with enhanced consistency across all three messaging platforms, and be more transparent about user data collection, data retention, and data sharing.
- Apply a minimum level of consistency in Community Standards across all three messaging platforms to facilitate improved user reporting.
- Consult with the Facebook 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.

- Apply the stricter standard in cases where separate content standards conflict (e.g., a message sent from WhatsApp to Messenger that violates Community Standards in the latter but not in the former).
- Develop publicly available, accessible, and understandable policy documents to disclose Meta's use of ML classifiers for detecting, flagging, and moderating accounts and content on messaging platforms.
- Examine whether and how ML classifiers for detecting, flagging, and moderating accounts and content on messaging platforms could result in discrimination.
- 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.
- Report the amount of problematic activity detected and accounts suspended on messaging platforms, as well as the success rates of the detection, disaggregated by relevant factors such as gender, geography, or age.
- 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.
- 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.
- Provide more information about how Meta's appeals process works in end-to-end encrypted platforms.
- Increase the speed and capacity of reporting and appeals processes, especially for vulnerable groups.
- Assess the grievance, reporting, and appeals process against the UNGPs effectiveness criteria for nonjudicial grievance mechanisms (i.e., legitimacy, accessibility, predictability, equitability, transparency, rights-compatible, source of continuous learning).
- Integrate human rights due diligence into privacy reviews and data protection assessment procedures.

Public Policy

Recommendations for how Meta should engage with key external stakeholders, such as law enforcement and civil society.

RECOMMENDATIONS

- 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.

- Engage policymakers about conflicting regulatory requirements that unnecessarily pit privacy rights against protecting users from broader harm, such as the European Privacy Directive.
- 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.
- Organize internal workshops and invite experts and academics who work on content-moderation techniques in an end-to-end encrypted environment to discuss the pros, cons, and feasibility of various mitigation techniques for specific issues.
- Continue to explore ways to provide data and other information for researchers focused on end-to-end encrypted messaging.
- 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.
- 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.
- Devote resources toward more accurately quantifying the scope of child sexual abuse material online and the corresponding harm to victims.
- Partner with children's rights organizations and educator groups to develop new children-specific training modules and tools tailored for the context of end-to-end encrypted messaging.
- Create issue-specific working groups within Meta's Safety Advisory Board and among its "trusted partners."
- Develop innovative methods to categorize reports and summarize their associated metadata for the National Center for Missing and Exploited Children (NCMEC).
- Continue to actively work with anti-trafficking organizations that have built relationships with survivor communities.
- Proactively collaborate with, train, and inform law enforcement about how to achieve their objectives in an end-to-end encrypted world in a rights-respecting way, such as by detecting and prosecuting crimes using alternative sources of digital evidence. This collaboration should be done on a case-by-case basis, based on the rule of law context of the jurisdiction involved, and have limited objectives to prevent misuse of new capabilities or related adverse human rights impacts.
- Continue working with other social media and internet companies to explore techniques to mitigate actual and potential human rights impacts of end-to-end encrypted messaging.
- Publicly communicate a strategy and action plan to address the adverse human rights impacts of end-to-end encrypted messaging, including progress toward achieving these recommendations over time.

Report Endnotes

1 A description of the UN Guiding Principles on Business and Human Rights "cause / contribute / directly linked" framework and its use in this assessment is found in Section 8.

2 https://www.facebook.com/notes/mark-zuckerberg/a-privacy-focused-vision-for-social-networking/10156700570096634/.

3 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.

4 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: https://scontent.whatsapp.net/v/t39.8562-34/271639644_1080641699441889_2201546141855802968_n.pdf/WhatsApp_Security_Whitepaper.pdf?ccb=1-5&_nc_sid=2fbf2a&_nc_ohc=m3hMa_fQzAIAX_hdEZb&_nc_ht=scontent.whatsapp.net&oh=01_AVwTMe-GmF0h9dxEkMuBM_-7XOE7Z7Uneiu5Z5svO0qV8Q&oe=622B9DBE.

5 Note that metadata associated with communications are not encrypted.

6 https://www.nytimes.com/2019/11/19/technology/end-to-end-encryption.html.

7 This is a description of how end-to-end encrypted messaging works. There is an ongoing debate about the precise definition of end-to-end encryption, which we reference in the full assessment.

8 https://freedomhouse.org/report/freedom-world/2020/leaderless-struggle-democracy.

9 https://freedomhouse.org/report/freedom-net/2021/global-drive-control-big-tech.

10 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. See the full report for details.

11 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%20encryption%20(,integrity%20and%20authenticity%20for%20users.

12 For example, in "Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems," the Center for Democracy & Technology (CDT) 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 designed recipients of this data.

13 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; 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; 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.

14 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

15 Server-side means that the computation takes place on a web server, whereas client-side means the computation takes place on the device.

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

17 https://www.npr.org/2021/08/25/1027397544/nso-group-pegasus-spyware-mobile-israel; https://citizenlab.ca/tag/nso-group/.

18 Note that the maximum group size on WhatsApp is 256 users.

19 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.

20 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/.

21 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/.

BSR®

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