April 2022

Meta's Approach to Safer Private Messaging on Messenger and Instagram Direct Messaging

Meta

TABLE OF CONTENTS

Introduction

Meta is committed to addressing issues of safety, security, and privacy in a holistic way that accounts for the constantly evolving dynamics of the digital age. When it comes to private messaging in Messenger and Instagram Direct, we believe that people should have secure, private places where they have clear control over who can communicate with them and confidence that no one else can access what they share. In March 2019, we presented our privacy-focused vision for social networking and announced our plans to expand end-to-end encryption (e2ee) to Messenger and Instagram Direct Messaging (DMs).[1] E2ee is increasingly the market reality and is already the standard for private messaging. Communications that are e2ee reinforce safety and security and have increasingly become the standard expectation of users for their preferred communications platforms. Our plans will build on the model employed by WhatsApp (which has been end-to-end encrypted by default since 2016) and focus on protecting the contents of people's private messages and calls with end-to-end encryption, while also using a combination of other signals to help keep our users safe and to prevent our services being misused to cause harm.

Since 2016, Meta has invested more than $16 billion and we now have more than 40,000 people focused on safety and security across our platforms. We've collaborated with experts around the world to design products, policies, tools and technologies that make it as difficult as possible for people to use Meta services to cause harm. However, to keep those who engage with our platform safe, we can't afford to stand still: as bad actors' methods, users' expectations, and technologies change, our safety strategy needs to evolve, too.

Messenger and Instagram DMs help billions of people stay connected to those who matter most to them. When they connect, they expect their conversations to be private and secure. Our goal is to provide people with the safest private messaging apps by helping protect people from abuse without weakening industry leading security protections, like e2ee. We focus on preventing abuse from happening, giving people controls to manage their experience, and responding to potential harm effectively. Protecting people on our apps requires constant iteration, so we regularly review our policies, update our features, and consult with experts.

We want people to have a trusted private space that's safe and secure, which is why we're taking our time to thoughtfully build and implement e2ee by default across Messenger and Instagram

[1] https://about.fb.com/news/2019/03/vision-for-social-networking/

DMs. E2ee is designed to protect people's private messages so that only the sender and recipient can access their messages. So, if you're sharing photos or banking details with family and friends, e2ee allows that sensitive information to be shared safely.

And while the vast majority of people use messaging services to connect with colleagues, friends, and loved ones, a small number of people will attempt to abuse them to do harm, including to young people. We have a responsibility to protect our users and that means setting a clear, thorough approach to safety. We also need to help protect people from abuse while maintaining the protections that come with encryption. People should have confidence in their privacy while feeling in control to avoid unwanted interactions and respond to abuse. Privacy and safety go hand-in-hand, and our goal is to provide people with the safest private messaging apps.

Our approach to help keep people safe when messaging through Messenger or Instagram focuses on:

- Understanding different types of potential harm and designing strategies for disrupting them.
- Working to prevent abuse from happening at source. We'll do this by:
    - Placing security at the forefront of our designs through (i) privacy and security protections and (ii) e2ee to prevent malicious actors from targeting our services;
    - Detecting and acting on suspicious patterns of activity; and
    - Deterring bad actors through transparent safety notices, robust reporting tools, and effective engagement with law enforcement.
- Giving people more controls to help them protect their experience on our apps.
- Responding quickly if harm occurs by:
    - Making it easy for people to respond to harm, including blocking other users and reporting harmful content or behavior from bad actors;
    - Enforcing our Community Standards[2] when we receive reports about a user or content; and

[2] https://transparency.fb.com/policies/community-standards/

    - Sharing relevant information with the National Center for Missing and Exploited Children (NCMEC) and law enforcement (in accordance with applicable law and our Terms of Service).
- Reviewing and continually evolving our approach.

This approach builds on the approach employed by WhatsApp,[3] a private messaging service that has been end-to-end encrypted by default since 2016, tailored to the distinct nature and features of Messenger and Instagram DMs.

We strongly believe that e2ee is critical to protecting people's security. Breaking the promise of e2ee - whether through backdoors or scanning of messages without the user's consent and control - directly impacts user safety.

The values of safety, privacy, and security are mutually reinforcing; we are committed to delivering on all of them as we move to e2ee as standard for Messenger and Instagram DMs. Our goal is to have the safest encrypted messaging service within the industry, and we are committed to our continued engagement with law enforcement and online safety, digital security, and human rights experts to keep people safe. Based on work to date, we are confident we will deliver that and exceed what other comparable encrypted messaging services do. We're also committed to continuing to invest as threats and technology constantly change and evolve.

Understanding Different Types of Harm

We take our role in keeping abuse off our services seriously, and we've dedicated significant resources to understanding the types of potential harm that could occur. That's why we developed standards for permitted uses of Messenger and Instagram DMs. Our Community Standards apply to users all around the world, and refer to the types of content that we believe are unacceptable for users to share, including on Messenger and Instagram DMs.

The harms covered by the standards (and the standards themselves) are based on feedback from our users and the advice of experts in fields such as technology, security, law enforcement, public safety, and human rights. To ensure everyone's voice is valued, we take great care to include different views and beliefs, especially from people and communities that might otherwise be overlooked or marginalized. The standards continue to be developed and refined in

[3] See e.g., https://faq.whatsapp.com/general/how-whatsapp-helps-fight-child-exploitation/?lang=en

consultation with global safety experts, including independent online safety organizations and experts.

We know each harm referred to Community Standards requires a tailored approach, as explained later in this paper. In Messenger and Instagram DMs, we balance privacy considerations and regulations, while taking a targeted 'prevent, control, and respond' approach to harm types including, but not limited to:

VIOLENCE AND CRIMINAL BEHAVIOR
- Violence and Incitement
- Dangerous Individuals and Organizations
- Coordinating Harm and Promoting Crime
- Restricted Goods and Services
- Fraud and Deception

SAFETY
- Suicide and Self-Injury
- Child Sexual Exploitation, Abuse, and Nudity
- Adult Sexual Exploitation
- Bullying and Harassment
- Human Exploitation
- Privacy Violations

OBJECTIONABLE CONTENT
- Hate Speech
- Violent and Graphic Content
- Adult Nudity and Sexual Activity
- Sexual Solicitation

Even within abuse types, there may be nuance, as our case study below shows. In an additional example, our policies on Child Exploitation, Abuse, and Nudity also clearly outline our robust approach to a range of child exploitative content. In particular, we do not allow:

- any content that threatens, depicts, praises, supports, provides instructions for, makes statements of intent, admits participation in, or shares links of the sexual exploitation of children;

- any content that solicits imagery of child sexual exploitation, or nude or sexualized images or videos of children;
- any content that constitutes or facilitates inappropriate interactions with children; and
- any content that attempts to exploit minors by coercing money, favors, or intimate imagery with threats to expose intimate imagery or information, or sharing, threatening, or stating an intent to share private sexual conversations.

Case Study: Developing an "Intent" Framework for Child Sexual Abuse Material (CSAM)

Understanding the possible or apparent intent of a sharer is important to developing effective interventions. For example, to be effective, the intervention Meta makes to stop those who share this imagery based on a sexual interest in children will be different from the action it takes to stop someone who shares this content in a poor (and still inappropriate and harmful to the victim) attempt to be funny.[4]

Research, such as the work of Former Federal Bureau of Investigation (FBI) Supervisory Special Agent Ken Lanning for NCMEC in 2010,[5] and Meta's own child safety investigative team's experiences, suggests that people who share these images are not a homogeneous group; they share this imagery for different reasons.[6]

Using an "intent taxonomy" developed with experts, including NCMEC, Meta has undertaken extensive review of Cybertips to understand the novelty (i.e., if the image is previously known) and severity of the child exploitation imagery being shared, as well as the intent behind the sharing -- all key information needed to assess risk, prioritize reports and develop new ways to reduce sharing of this content. Currently, the majority of volume in Cybertips is the same content being reshared at scale. Meta has found that more than 90% of content reported to NCMEC on Facebook and Instagram was the same as or visually similar to previously reported

[4] *See* Malia Andrus, John Buckley, Chris Williams, *Understanding the intentions of Child Sexual Abuse Material (CSAM) sharer*s, Meta Research Blog (Feb. 23, 2021), https://research.facebook.com/blog/2021/02/understanding-the-intentions-of-child-sexual-abuse-material-csam-sharers/.

[5] Kenneth V. Lanning, *Child Molesters: A Behavioral Analysis*, National Center for Missing & Exploited Children (2010), https://www.missingkids.org/content/dam/missingkids/pdfs/publications/nc70.pdf.

[6] *See* Malia Andrus, John Buckley, Chris Williams, *Understanding the intentions of Child Sexual Abuse Material (CSAM) sharers*, Meta Research Blog (Feb. 23, 2021), https://research.facebook.com/blog/2021/02/understanding-the-intentions-of-child-sexual-abuse-material-csam-sharers/; Kenneth V. Lanning, *Child Molesters: A Behavioral Analysis*, National Center for Missing & Exploited Children (2010), https://www.missingkids.org/content/dam/missingkids/pdfs/publications/nc70.pdf.

content in the same period. In other words, a small number of images represent the large majority of the images shared (and reshared) and reported. Initial review has discovered that a very large portion of this content is shared without "malicious intent" - meaning it is not shared by people with a sexual interest in children.[7] While the sharing of this content is still harmful, these users are not likely to be the focus of law enforcement investigations.[8]

Preventing Harm at Source

Preventing harm of all types from happening in the first place is the best way to keep people safe. Our investment in prevention draws on a growing body of research that has recognized the effectiveness of crime and harm prevention. That's why we are working to prevent harm and abuse from happening at source.

Key to prevention is placing security at the forefront of our designs through strong default privacy protections and investing in default e2ee to prevent malicious actors from targeting our services.

As the section below shows, prevention is also at the core of the work Meta does to protect safety. When e2ee is standard, Meta will continue to disrupt harm related to Messenger and Instagram DMs using similar technology to that used to detect spam and scams. Without accessing/scanning the contents of our users' private messages (unless reported), we will identify suspicious behavior, then restrict account features to make it harder for those users to find and contact people they don't know, including children, disrupting potential harm before it happens.

Our product designs will help divert and deter would-be offenders, limit harmful interactions on our messaging services, and empower users to report suspicious behavior while educating them

[7] *See* Malia Andrus, John Buckley, Chris Williams, *Understanding the intentions of Child Sexual Abuse Material (CSAM) sharers*, Meta Research Blog (Feb. 23, 2021), https://research.facebook.com/blog/2021/02/understanding-the-intentions-of-child-sexual-abuse-material-csam-sharers/; Kenneth V. Lanning, *Child Molesters: A Behavioral Analysis*, National Center for Missing & Exploited Children (2010), https://www.missingkids.org/content/dam/missingkids/pdfs/publications/nc70.pdf.

[8] Kenneth V. Lanning, *Child Molesters: A Behavioral Analysis*, National Center for Missing & Exploited Children (2010), https://www.missingkids.org/content/dam/missingkids/pdfs/publications/nc70.pdf.

on how to avoid harmful interactions. However, there is no "one size fits all" response to harm, and no one solution for any harm type, or even any subset of a harm type. All require a nuanced understanding and approach.

We'll also continue to invest in our industry-leading systems, tools, and strategies to detect and act on suspicious patterns of activity. And we are increasingly focused on using upstream detection methods, including disrupting entire networks of bad actors before they can use messaging to cause harm in the first place. We will continue to invest in our ability to detect harmful behavioral patterns using non-content signals, content on non-encrypted surfaces like Facebook and Instagram, and user reports of messaging content to identify and respond to potential abuse. With default e2ee, users will continue to have robust tools to report abusive and harmful content, enabling, for example, prioritization of Cybertips for CSAM and prevention of worse or ongoing harms. These efforts will further support effective engagement and response to law enforcement to prevent real-world harm. When it comes to protecting children and addressing the sharing of CSAM, our approach goes beyond a "detect, report, and remove" model.

1. Preventing harm through e2ee

E2ee is an important component of safety, particularly when we focus on prevention. E2ee protects people from serious and common crimes like hacking and identity theft and enables secure communications for individuals who may be targeted by authoritarian or illiberal regimes, as well as people who may be subjected to domestic violence and abuse or hate crimes.

In addition to protecting our digital systems from government or other external intrusion, e2ee serves as an effective measure to prevent improper use of communications by malicious actors with access to a platform's own systems. That would include both employees who seek to abuse their legitimate access[9] for nefarious ends, as well as threat actors or criminals seeking to gain unauthorized access to the platform's systems.[10]

---

[9] *See, e.g.*, Joseph Cox, *Leaked Document Says Google Fired Dozens of Employees for Data Misuse*, Vice (Aug. 4, 2021), https://www.vice.com/en/article/g5gk73/google-fired-dozens-for-data-misuse; Joseph Cox, *Snapchat Employees Abused Data Access to Spy on Users*, Vice (May 23, 2019), https://www.vice.com/en/article/xwnva7/snapchat-employees-abused-data-access-spy-on-users-snapli on.

[10] *See, e.g.*, Brian Fung, *Twitter Hackers Accessed Direct Messages of 36 Accounts, Company Says*, CNN Business (July 22, 2020),

We believe e2ee and promoting people's safety go hand in hand. Indeed, e2ee is an important tool for protecting the right to privacy for users around the world. In a digital age, giving users control over who has access to their data is fundamental to the concept of privacy. Many of the most sensitive conversations are now conducted via digital services - conversations with doctors, lawyers, counselors, partners, children, friends, and co-workers. It is vital that people have a means to prevent unintended third parties from viewing their private conversations. As Professor Ciaran Martin, former head of cybersecurity at the United Kingdom's Government Communication Headquarters (GCHQ), notes:

> End-to-end encryption exists, it works, and it makes sense. Tech companies know it and privacy campaigners know it. But so too do citizens. And, frankly, so too do policymakers.[11]

For years, academics, researchers, and many government officials have agreed that e2ee is the best technology currently available for protecting the privacy and security of sensitive information and communications. For example, in a November 2020 statement,[12] nine NGOs, including Privacy International and Article 19, wrote that "end-to-end encryption in particular, provides a guarantee that our private communications and information will be secure, and not vulnerable to being hacked or otherwise accessed without our consent," and that it is "a technology that millions of people across the world rely upon for their privacy, safety and security," including "journalists, human rights defenders, whistle-blowers, activists, and minorities vulnerable to persecution." Inherent to end-to-end encryption are fundamental human rights like freedom of expression, freedom of information, and association and assembly, and encryption directly impacts the public's right to information by allowing investigative journalists to guarantee source protection. The authors emphasized that encryption is recognized by major human rights bodies around the world, including UN Special Rapporteurs, the UN Human Rights Council, and the Freedom Online Coalition. For example, in his capstone report on the role encryption plays in free expression, former UN Special Rapporteur for

---

https://www.cnn.com/2020/07/22/tech/twitter-hack-direct-messages/index.html.

[11] *End-to-End Encryption: The (Fruitless?) Search for a Compromise*, Bingham Centre for the Rule of Law (Nov. 2021), https://www.bsg.ox.ac.uk/sites/default/files/2021-11/End-to-end%20Encryption%20Ciaran%20Martin%20Blavatnik%20School.pdf.

[12] *Joint Civil Society Statement on Encryption*, Article 19 (Nov. 13, 2020), https://www.article19.org/resources/uk-joint-civil-society-statement-on-encryption/; *see also UK: Joint Letter to MPs: End-to-End Encryption Keeps Us Safe*, Article 19 (June 14, 2021), https://www.article19.org/resources/uk-joint-letter-to-mps-end-to-end-encryption-keeps-us-safe/.

Freedom of Expression and Opinion David Kaye affirmed that "encryption and anonymity, and the security concepts behind them, provide the privacy and security necessary for the exercise of the right to freedom of opinion and expression in the digital age." "Such security may be essential for the exercise of other rights," the report concluded, "including economic rights, privacy, due process, freedom of peaceful assembly and association, and the right to life and bodily integrity."[13]

For example, whilst end-to-end encrypted chats are already available as an option on Messenger and are by default on WhatsApp,[14] we've made encrypted one-to-one chats available on Instagram for all adults in Ukraine and Russia to protect their information from illicit use, including by the invading army. We'll also show notifications at the top of people's direct message inboxes to let them know they can switch to an encrypted conversation if they want to.

Importantly, as affirmed in Business for Social Responsibility's (BSR) Human Rights Impact Assessment (HRIA), the privacy protections of Meta's e2ee messaging platforms "keep people safe from bad actors who would use their message content to cause them bodily harm or detain them arbitrarily."[15] The HRIA acknowledges 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" means that vulnerable groups in particular are "dependent on the right to privacy to enable these other rights."[16] These groups 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 LGBTQI+ rights groups, among others.[17] While the HRIA acknowledges the possible risk of use of e2ee to facilitate the trafficking of adults and children, to share CSAM, or to plan terrorist attacks[18] unless these risks are mitigated (in many of the ways we explain in this paper), it ultimately concludes that "[e]xtending end-to-end encryption across messaging platforms will provide vital safety protections for vulnerable users and other rightsholders around the world."[19]

---

[13] https://documents-dds-ny.un.org/doc/UNDOC/GEN/G15/095/85/PDF/G1509585.pdf?OpenElement
[14] This update is not available to business accounts on Instagram.
[15] BSR, 2022. "Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption" at 34.
[16] BSR, 2022. "Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption" at 63.
[17] BSR, 2022. "Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption" at 63.
[18] BSR, 2022 "Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption" at 33-35.
[19] BSR, 2022 "Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption" at 44.

2. Detecting patterns of harm and abuse

We are committed to using technology-driven solutions across the data that we have available to us and are permitted to use for this purpose, to detect behavioral patterns of abuse. For example, Meta's Machine Learning (ML) tools focus on early detection and prevention. Our ML technology currently looks across the public facing surfaces on our family of apps - like account information and photos uploaded to public spaces like Facebook and Instagram - to detect suspicious activity and abuse before it reaches messaging. When e2ee is default, we will also use artificial intelligence, subject to applicable law,[20] to proactively detect accounts engaged in malicious patterns of behavior instead of scanning private messages.

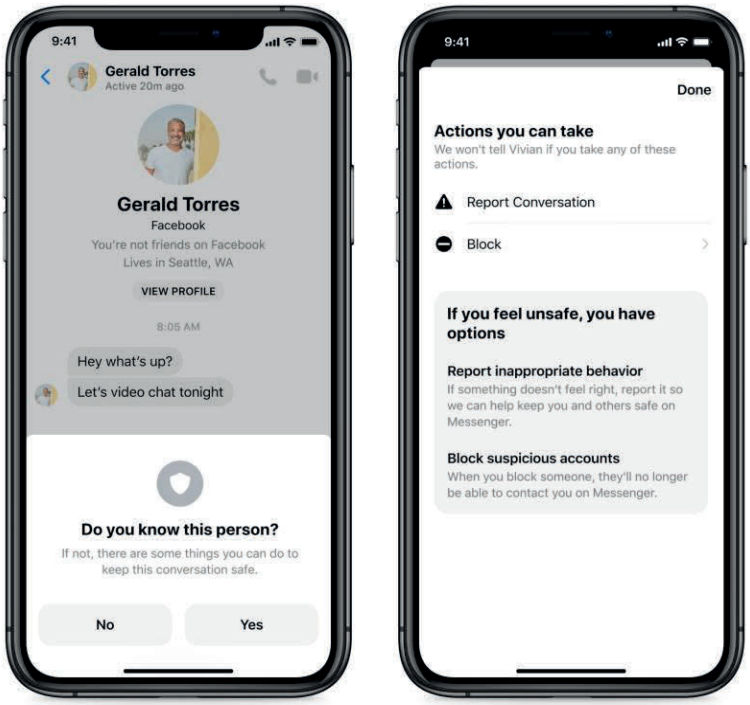
Investigations have shown that bad actors often reveal their intentions with obvious public signals. Some of these examples within the context of child sexual abuse include friending accounts with clear child-sexualizing content, using coded language in bios, searching for egregious terms, or joining questionable groups. Notably, key signals of harm involve non-encrypted content that will remain available when our messaging services move to default e2ee.

Based on these signals and others, we can take action. This includes prevention actions such as limiting their ability to interact with others and engage in offending behavior (e.g., enhanced user education), or environmental changes (e.g., removing the friend button or preventing a suspected account from messaging with a minor).

Much like email spam filters, analyzing behavioral signals in a private space with privacy-preserving techniques provides opportunities to detect bad actors connecting with one another, and most importantly, to detect when they may be targeting victims. For example, if an adult repeatedly sets up new account profiles and tries to connect with minors they don't know or messages a large number of strangers, we can intervene by taking actions such as blocking

[20] https://about.fb.com/news/2020/12/changes-to-facebook-messaging-services-in-europe/

them from interacting with minors in order to prevent inappropriate interactions.



We have found through research that when bad actors are simply blocked from connecting with or messaging vulnerable users, they will simply keep looking for new ways to offend. Instead, Meta focuses on limiting bad actors' capabilities within the product experience (e.g., removing friend button or search options in Facebook) to ensure they cannot use Meta apps to facilitate harm (similar to the redirect method used in counter-terrorism) and redirecting them to educational resources.[21] For actors who repeatedly attempt to contact children or others for harmful purposes, we can take more serious actions including disabling and removing their accounts.

3. Empowering Messenger and Instagram users through built-in prevention and education

We also aim to prevent harm by empowering and educating users on how to identify and protect themselves from unwanted interactions on Messenger and Instagram DMs through in-app safety notices and easy to use reporting tools.

[21] https://redirectmethod.org/.

A key part of our education strategy is to provide users, especially young people, with in-app advice on avoiding unwanted interactions. For example, we provide educational signposting to all users on Facebook that they should only accept friend requests from people they know. If a user blocks, or deletes a connection request, it is likely that the user does not want to interact with the requester. This prompts us to ask if the user wants to report.

We've also seen tremendous success with our safety notices on Messenger, which are banners that provide tips on spotting suspicious activity and taking action to block, report, or ignore/restrict someone when something doesn't seem right.[22] We developed these safety tips using machine learning to help people avoid scams, spot impersonations and, most urgently, flag suspicious adults attempting to connect to minors. In November 2021 alone, more than 100 million people saw safety notice banners on Messenger. And, importantly, this feature works with e2ee.

Our Terms prohibit children below the age of 13 from opening an account on Facebook, Instagram, and Messenger. For those young users who are old enough to use our apps, we have deployed numerous default protections, policies, and tools to prioritize their safety. For example:

- Adults cannot initiate Messenger chats and Instagram DMs with minors they are not connected to.
- On Messenger, we have Safety Notices alerting teens of suspicious activity and prompting them to take action to block, report, or ignore/restrict someone when something doesn't seem right.
- People under 18 years also do not have access to certain services supported by Messenger such as Marketplace, Mentorships, Fundraisers, Dating, and Blood Donation.
- We protect certain information such as contact info, school, or birthday from appearing publicly.
- Location sharing is off for all users by default. When a teen turns on location sharing, we include a consistent indicator as a reminder that they're sharing their location.
    - On Instagram, for new users under 18 or under 16 (depending on country), we set their accounts to private by default, which means that their content cannot be seen by others without permission of the user.

22 https://messengernews.fb.com/2020/05/21/preventing-unwanted-contacts-and-scams-in-messenger/

- We provide in-app notifications to encourage Instagram users under 18 or under 16 (depending on country) to use their privacy settings and to educate them about the consequences of having accounts made public.
    - We prevent the profiles of people under 18 years old from appearing in search tools outside of Facebook.
- On Facebook, new teen accounts are automatically defaulted to share their posts with "friends" only.
- We remove teens from the "suggested friends" of potentially suspicious adults and we remove the ability for certain adults to friend minors.

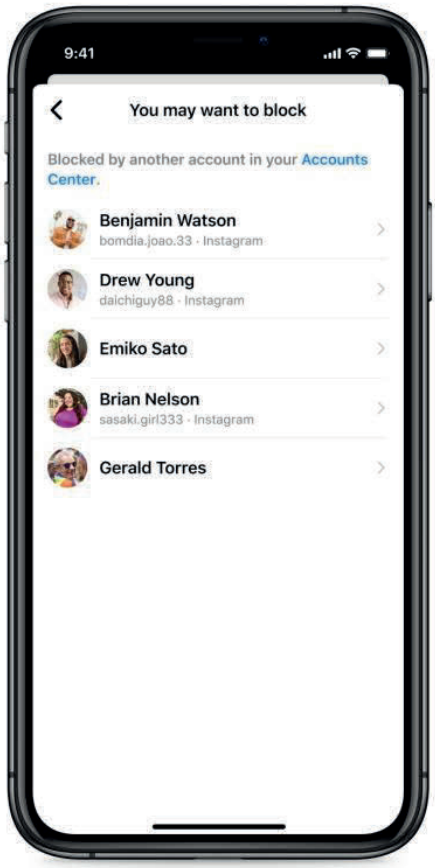
We also want to educate more people to act if they see something and avoid sharing harmful content, even in outrage. We have begun sending alerts informing people about the harm that sharing child exploitation content can cause, even when done in outrage or to raise awareness, by warning them that it's against our policies and will have legal consequences. We've also launched a global "Report it, Don't Share it" campaign reminding people of the harm caused by sharing this content and the importance of reporting this content.

Giving People More Choice and Control

While we put in place strong default security and privacy protections, we know that our users are diverse, covering all age ranges and countries and with different requirements from their messaging experiences. Emerging creators often want increased reach to potential followers, while other people want tight-knit circles. In addition to our efforts to prevent harm, we are giving users more controls of their messaging inbox to account for the variety of experiences people want.

Over the past few years, we've improved the options for reviewing chat requests and recently built delivery controls that let people choose who can message their chats list, who goes to their requests folder and who can't contact them at all. To help people review these requests in the safest way possible, we blur images and videos, block links, and let people delete requests to chat in bulk. (*Note: some features may not be available to everyone*.)[23]

[23] https://about.fb.com/news/2020/12/changes-to-facebook-messaging-services-in-europe/.



People can already block unwanted contacts in Messenger, so we're introducing the ability to block unwanted contacts seamlessly across Instagram DMs and Messenger. We are making it easy to block contacts from strangers-- a feature that is switched on by default for any user we identify as a potential minor.

We also recently announced Hidden Words on Instagram so people can determine for themselves what offensive words, phrases and emojis they want to filter into a Hidden Folder.[24] The user decides on a list of potentially offensive words, hashtags, and emojis by default, even if they don't break our rules. This is also part of our effort to take a broader approach to safety, and will work on e2ee on-device, under the control of the user.

24 https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse.

Responding Quickly to Potential Harm

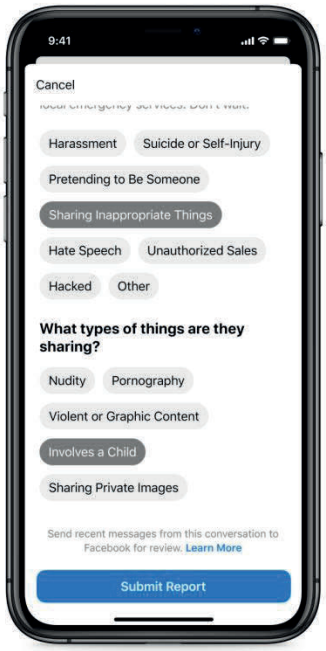
When we become aware of potential abuse on our services, we respond quickly to mitigate any harm. We do this by making it easy for people to report harmful content, enforcing our Community Standards when we receive reports about users or content, and sharing data with NCMEC and our law enforcement agencies.

1. User reporting

Reporting is an essential tool for people to stay safe and help us respond to abuse effectively. We encourage our users to report content to us that they believe violates our policies using the dedicated tools we have designed for our services. This includes Pages, Groups, profiles, individual posts, and comments on Facebook and Instagram and accounts and chats on Messenger and Instagram DMs.

On our messaging services, we're making it much easier to report harm and educating people on how to spot scammers and impersonators by redesigning our reporting feature to be more prominent in Messenger and Instagram DMs. We also recently made it easier to report content for violating our child exploitation policies.[25] When reporting harm, people can select "involves a child" as an option, which, in addition to other factors, prioritizes the report for review and action. Our goal is to encourage significantly more reporting by making it more accessible, especially among young people. As a result, we're seeing close to 50% year-over-year growth in reporting, and we're taking action to keep Messenger and Instagram DMs safe.

[25] https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/



We'll continue to enforce our Community Standards on Messenger and Instagram DMs in a default e2ee environment. When users choose to report and send us data from their device, we'll be able to see and review up to 30 of the most recent messages sent in that conversation. Messages are decrypted on the user's device and securely sent to Meta via the user's reporting action.[26] This allows us to take action if violations are detected — whether they are scams, bullying, harassment, violent crimes, or child exploitation.

Meta also encourages user reporting based on non-content signals. In addition to the safety notices described already, we are also conducting research to identify additional opportunities to not only offer reporting but prompt it and identify how to most effectively do so, especially with regard to minors. For example, we are exploring the effectiveness of prompting reporting when someone blocks a user or deletes a thread. Early results suggest this may be a very promising direction that can make a significant difference to user reporting.

2. Working with law enforcement

Meta works to ensure safety in our community, online and offline, including by working with law enforcement. In addition to empowering and educating people to use our safety and reporting tools, we engage with law enforcement agencies to respond to valid legal requests and may

[26] https://www.facebook.com/help/786613221989782.

provide information to law enforcement that will help them respond to emergencies, including those that involve the risk of immediate harm, suicide prevention and the recovery of missing children - all consistent with applicable law, our Terms of Service and data policy, and human rights/international standards. We scrutinize every government request we receive to make sure it's legally valid and, when we comply, we produce narrowly tailored information to respond to that request.

As part of our ongoing effort to share more information about the requests we have received from governments around the world, Meta regularly produces reports on 'Government Requests for User Data' in its Transparency Center,[27] to provide information on the nature and extent of these requests and the strict policies and processes we have in place to handle them. Our website also provides further details, including operational guidelines for law enforcement officials seeking records from Meta ("Law Enforcement Guidelines")[28] and our dedicated Law Enforcement Online Request System (LEORS).

We value the work of law enforcement agencies around the world and share the goals of keeping people safe. Meta has a strong history of engagement with law enforcement agencies on critical safety issues. This applies across our services - including for Messenger and Instagram DMs - and will not change following default e2ee.

In response to valid legal requests or where there is an imminent risk of harm to a child or risk of death or serious physical injury to any person, Meta will still be able to produce available data that can support law enforcement investigations. In fact, Europol's annual digital evidence report found 85% of law enforcement surveyed cited Basic Subscriber Information (BSI) (including email address and phone number) and traffic data (IP address) as the types of data needed most often in investigations.[29] Both of these types of data will remain available to Meta and can be produced to law enforcement when our messaging services move to e2ee by default.

This can be vital information to law enforcement when responding to emergencies, including helping law enforcement to locate people at imminent risk of physical harm, for suicide prevention, and for the recovery of missing children.

---

[27] Meta Transparency Center, "Government Requests for User Data", https://transparency.fb.com/data/government-data-requests/

[28] Facebook Safety Center, "Information for Law Enforcement Authorities" (last visited Dec. 28, 2021), https://www.facebook.com/safety/groups/law/guidelines/; Facebook Safety Center, "Facebook and Law Enforcement" (last visited Dec. 28, 2021), https://www.facebook.com/safety/groups/law.

[29] SIRIUS EU Digital Evidence Situation Report (2020), at 13, available at https://www.europol.europa.eu/cms/sites/default/files/documents/sirius_desr_2020.pdf.

Meta's messaging services also sit alongside its other products - including the main Facebook and Instagram social networking platforms. Through appropriate legal process, law enforcement may request available data that reflects a user's activity across our private and public-facing surfaces - with the potential to provide law enforcement with a better understanding of how users may be abusing our services to perpetrate harm.

It is also a mistake to consider the information that Meta can provide in isolation, as that may only be part of the puzzle and the ultimate picture that law enforcement is able to build from multiple sources. For example, in a default e2ee environment users will still maintain access to their messaging content and law enforcement may still be able to obtain this content directly from users or their devices.

3. Partnering with NCMEC

In addition to engaging with law enforcement, we partner closely with NCMEC to mitigate the spread of CSAM on the internet. We report all apparent instances of child sexual exploitation appearing on our platforms from anywhere in the world to NCMEC,[30] including content brought to our attention by government requests. NCMEC coordinates with law enforcement authorities from around the world to help children. If a request relates to a child exploitation or safety matter, law enforcement may specify those circumstances (and include relevant NCMEC report identifiers) in the request to ensure that Meta is able to address these matters expeditiously and effectively.[31]

We will continue to detect suspicious activity on public surfaces and on non-messaging platforms, including users' Facebook and Instagram activity. For instance, Meta can still scan for images on public and non-encrypted surfaces when e2ee is standard. If the information from those scans amounts to facts and circumstances of CSAM, Meta will continue to report it.

In child exploitation cases involving Messenger or Instagram DMs, we'll continue to report violating accounts to NCMEC.[32] We're able to share data like account information, account activity, content from non-encrypted parts of our services (such as profile photos) and inbox content from user reported messages to determine compliance with our Terms of Service and

[30] US law requires service providers with functions in the US to report CSAM to NCMEC and, in doing so, the reporting of the content to NCMEC does not amount to illegal distribution of CSAM.
[31] Facebook Safety Center, "Information for Law Enforcement Authorities" (last visited Dec. 28, 2021), https://www.facebook.com/safety/groups/law/guidelines/.
[32] https://transparency.fb.com/policies/improving/working-with-law-enforcement/

Community Standards. We'll continue to iterate on this approach. We have been making continuous improvements to our detection systems that do not rely on proactive scanning of inbox content.

When producing reports to NCMEC, we want to ensure that the information we provide are actionable by law enforcement to support people's safety, security, and privacy.

Reviewing and Continually Evolving our Approach

As technology is constantly evolving and bad actors change their techniques, preventing abuse on our apps therefore requires constant iteration. We regularly review our policies and features, listen to feedback from experts and people using our apps, including Messenger and Instagram DMs, to stay ahead of people who may not have the best intentions.

While building a trusted space requires ongoing innovation, flexibility, creativity, and engagement with outside experts, we believe that this approach of prevention, control, and response offers a framework to get people the protection they need and deserve. Privacy and safety go hand-in-hand, and we're committed to making sure they are integral to people's messaging experiences.

It is a top priority of ours to continually improve our ability to prevent, disrupt, and respond to bad actors who misuse any of our apps, including Messenger and Instagram DMs, to facilitate harm. Our efforts to improve mean we support a collaborative, rights-respecting, solutions-based approach that works across stakeholders for two reasons: to build better products for users, including to prevent harm on our apps, and to make our reports actionable to stop bad actors from returning and continuing to perpetrate harm. For example, while early testing on Meta's intent-based content and non-content signals models is promising, additional data from law enforcement on the investigations they open and cases they prosecute based on the information we provide would help further this work.

We are open to working with governments and stakeholders on industry-wide success metrics that better assess progress on reducing harm in messaging, including child sexual exploitation.

Working Across Global Stakeholders
Child protection requires a global and comprehensive response from industry, law enforcement, government, civil society, and families, which is why Meta is committed to working with child safety stakeholders worldwide to build and support the child safety

ecosystem. Because online child exploitation is a global internet problem, it demands a global internet solution.

Meta has worked with the European Commission for more than a decade, as a member and signatory to the CEO Coalition to make a better Internet for kids and its subsequent five point action plan.

We remain a member of the European Commission's Alliance to Better Protect Minors Online and actively participate annually in Safer Internet Day celebrations, as well as the Safer Internet Forum.

Several organizations and initiatives bring together industry and other players in the fight against child sexual abuse such as the Technology Coalition, an association dedicated solely to eradicating the sexual exploitation of children online, the ICT Coalition, an industry-led body that encourages cross-industry collaboration in support of the EU's child safety goals, and international multi-stakeholder organizations like the WePROTECT Global Alliance to end child exploitation.

In 2020, Meta, Google, Microsoft, and 15 other technology companies came together to announce Project Protect: A plan to combat online child sexual abuse – a renewed commitment and investment expanding the Technology Coalition's scope and impact to protect children online and guide its work for the next 15 years.

We are also proud of two recent collaborations with the governments of Australia, Canada, New Zealand, the United Kingdom, and the United States: 1) we partnered with End Violence Against Children, Microsoft, Google, Twitter, Roblox, and Snapchat, to develop a new campaign, 'Stay Safe at Home, Stay Safe Online,' which provides safety tips for parents, caregivers, and children; and 2) the Voluntary Principles to Counter Online Child Sexual Exploitation and Abuse, which sets out a framework of 11 actions as part of tech firms' fight against online child sexual abuse.

Additionally, we work closely with our Safety Advisory Board of leading online safety non-profits, as well as over 400 safety experts and NGOs from around the world, including specialists in the area of child sexual exploitation and victim support. We are committed to educating people on how to stay safe online and work with industry, NGOs, and other stakeholders to ensure people have the resources they need to stay safe.

In 2019, we launched Stop Sextortion, a dedicated hub in our Safety Center developed by Thorn, a leading NGO in the fight against child sexual abuse, with resources for teens, caregivers and educators seeking support and information related to sextortion.

In June 2020, we published a TTC Labs Youth Design Guide, co-designed with young people, as part of the current European Commission' Youth Pledge. In addition, we continue to deploy Facebook's Get Digital in a number of EU Member States to bring digital citizenship into classrooms.

As part of our prevent, detect, and respond strategy, we have recently partnered and invested in three emerging online safety challenges:

- **Self-generated teen nudity**

  In the last two years, youth self-generated CSAM has emerged as the predominant type of content reported via hotlines in many European countries. In 2020, Meta, in partnership with the UK's Internet Watch Foundation and NSPCC, launched Report Remove, a self-referral channel, with a safeguarding age verification and support service for young people, without fear of criminalization. This pilot is being expanded outside of the UK and has provided important lessons on how to better tackle this new and concerning trend.

- **CSAM Offender Diversion**

  In partnership with the Institute of Sexology & Sexual Medicine of Charité - Universitätsmedizin Berlin, as well as a range of other organizations including the Lucy Faithfull Foundation, Meta is exploring opportunities to better support people who feel attracted to, or sexually aroused by, children and adolescents, through the Troubled Desire project. This project offers an online self-management tool in more than 10 languages for individuals with a sexual attraction to minors who don't have the chance to get in contact with therapists.

- **Non-malicious sharing and distribution of CSAM**

  In June 2021, we launched a PSA (video) campaign in 14 countries, in partnership with local child safety organizations to remind people that sharing child exploitative material, even in the context of outrage or condemnation, causes further harm to the child and is illegal. The key message is: "Report it. Don't Share it." This PSA comes on the heels of recent research we conducted on our CyberTips to NCMEC that found a majority of the

> reports we make to NCMEC are re-shares; a small handful of countries are responsible for these reports and people mostly share this content out of outrage and not because they have sexual interest in children. The campaign targets top sharers/receiver countries.[33]

We remain deeply committed to combating child exploitation and abuse across our services and around the world and will continue to seek out opportunities to contribute and advance multi-stakeholder efforts to eradicate its presence online.

Why Weakening Encryption Impacts User Safety and Security

In October 2020, a UNICEF working paper concluded that "[e]nd-to-end encryption is necessary to protect the privacy and security of all people using digital communication channels" including "children, minority groups, dissidents and vulnerable communities." The working paper also noted that the UN Special Rapporteur on Freedom of Expression "has referred to end-to-end encryption as 'the most basic building block' for security on digital messaging apps, as well as being important for national security."[34]

Experts in cryptography and computer science are in near unanimous agreement - and have been for many years - that strong encryption is the best defense against vulnerabilities or weaknesses in our digital systems. This model is threatened if we allow for "backdoors" or so-called "exceptional access" that weakens the security of e2ee systems, and which experts agree will inevitably be discovered and sought to be exploited by malicious actors on a much larger scale. For example, fifteen leading experts concluded in a seminal paper in 2015 that proposals to design such "exceptional access" into encrypted systems "are unworkable in practice, raise enormous legal and ethical questions, and would undo progress on security at a time when Internet vulnerabilities are causing extreme economic harm."[35]

---

[33] https://research.facebook.com/blog/2021/02/understanding-the-intentions-of-child-sexual-abuse-material-csam-sharers/

[34] Daniel Kardefelt-Winther et al., *Encryption, Privacy and Children's Right to Protection from Harm*, UNICEF Office of Research - Innocenti, WP-2020-14, at 3 (Oct. 2020), https://www.unicef-irc.org/publications/pdf/Encryption_privacy_and_children%E2%80%99s_right_to_protection_from_harm.pdf.

[35] Harold Abelson et al., *Keys Under Doormats: Mandating Insecurity by Requiring Government Access to All Data and Communications*, MIT Computer Science and Artificial Intelligence Laboratory, at 1 (July 6, 2015), http://dspace.mit.edu/bitstream/handle/1721.1/97690/MIT-CSAIL-TR-2015-026.pdf.

This report, entitled "Keys Under Doormats," identified three primary problems with designing exceptional access into systems. First, designing for exceptional access is incompatible with security best practices that are used to secure communications. Second, such exceptional access mechanisms substantially increase the complexity of a system, and "complexity is the enemy of security" because "every new feature can interact with others to create vulnerabilities." And third, exceptional access creates "concentrated targets" that would become attractive to bad actors - by compromising whatever system retains decryption credentials, a bad actor would be able to compromise every communication to which the exceptional access mechanism allows access.[36]

Another threat to the security provided by e2ee comes in the form of various new proposals for "technical solutions," to proactively detect or monitor content on e2ee messaging platforms. Such tools are generally referred to as "client-side scanning." Client-side scanning would involve leveraging a user's device to scan for the presence of certain harmful or prohibited content, such as known CSAM, and report any content that matches a known database to third parties – like the provider itself or to law enforcement – without the users' consent, control or knowledge. Following years of debate, including legislative and technical proposals,[37] many of the same security and cryptography experts that had written the "Keys Under Doormats" paper argued last year against this kind of proactive monitoring of encrypted messages – in this case, for CSAM. The authors of this paper, entitled "Bugs in Our Pockets," argued that client-side scanning "creates serious security and privacy risks for all society while the assistance it can provide for law enforcement is at best problematic."[38]

The report in turn addressed five main risks that new proposals for client-side scanning present, namely that: 1) the move toward scanning hardware for unshared content is a new frontier in the invasion of privacy, including surveillance and censorship; 2) client-side scanning increases the attack surface for cybersecurity risks to all devices where such scanning takes place; 3) the

---

[36] *Id.* at 2-3.
[37] *See, e.g.*, https://www.eff.org/deeplinks/2022/02/its-back-senators-want-earn-it-bill-scan-all-online-messages; https://www.eff.org/deeplinks/2020/03/earn-it-bill-governments-not-so-secret-plan-scan-every-message-online; https://edri.org/our-work/a-beginners-guide-to-eu-rules-on-scanning-private-communications-part-1/; https://homeofficemedia.blog.gov.uk/2021/09/08/new-safety-tech-fund-challenge/; https://techcrunch.com/2021/09/03/apple-csam-detection-delayed.
[38] Harold Abelson et al., *Bugs in our Pockets: The Risks of Client-Side Scanning*, at 1 (October 15, 2021), https://arxiv.org/pdf/2110.07450.pdf.

technology itself is not effective at achieving its intended objective, due to successful evasion attacks, false positive/negatives, and fallible algorithms; 4) on a technological level, client-side scanning is currently not practicable or implementable at scale across devices with varying computational capacity (and will remain so notwithstanding future speculative technologies); and 5) in many jurisdictions such scanning may present legal and/or constitutional challenges.

Meta believes that any form of client-side scanning that exposes information about the content of a message without the consent and control of the sender or intended recipients is fundamentally incompatible with user expectations of an e2ee messaging service. People that use e2ee messaging services rely on a basic promise: that only the sender and intended recipients of a message can know or infer the contents of that message.

Both these threats to e2ee – the use of exceptional access as well as as client-side scanning – are further addressed in BSR's HRIA, which explained that any benefits associated with these tools can be "undermined in scenarios where client-side scanning is abused, weakens end-to-end encryption, or leads to a regulatory slippery slope."[39] The HRIA emphasizes that security and cryptography experts have raised concerns about the technical integrity of proposals for deploying client-side scanning systems since there is a real "risk that bad actors may take advantage of the technical vulnerabilities of these solutions to game the system."[40] Tackling harmful content requires a collaborative, ongoing effort between civil society, the technology community, and law enforcement agencies, but it need not require pursuing options that could result in significant human rights impacts on privacy, freedom of expression, and the physical safety of particularly vulnerable groups.[41]

In 2016, leading security expert Bruce Schneier wrote that "[m]any technological security failures of today can be traced to failures of encryption," referencing the U.S. Office of Personnel Management breach as well as a variety of commercial data thefts.[42] He argued that "[a]dding backdoors will only exacerbate the risks" because it is impossible for technologists to build an access mechanism "that only works for people of a certain citizenship, or with a particular morality, or only in the presence of a specified legal document." Any such mechanism can be

[39] BSR, 2022. "Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption" at 87.
[40] BSR, 2022. "Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption" at 86.
[41] BSR, 2022. "Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption" at 88-89.
[42] Bruce Schneier, *Security vs. Surveillance*, Schneier on Security (Feb. 1, 2016), https://www.schneier.com/essays/archives/2016/02/security_vs_surveill.html.

exploited; as Schneier wrote again in 2018, "Demanding that technology companies add backdoors to computers and communications systems puts us all at risk."[43]

The state of the world has not changed in any way that would undermine these sentiments in the ensuing years - in fact, cyber threats have increased dramatically in quantity and severity, and the increasing involvement of nation-states has contributed significantly to the rising risks. The recent revelations regarding the widespread misuse of the Pegasus spyware tool to access the devices and the private communications of activists, journalists, and political leaders provides a chilling example of the current threat landscape.[44] History has shown that even the most sensitive and highly secured systems can be breached, even if they are not connected to the internet.[45]

As more than 100 advocacy groups wrote in an open letter to Meta in 2019, "Given the remarkable reach of Facebook's messaging services, ensuring default end-to-end security will provide a substantial boon to worldwide communications freedom, to public safety, and to democratic values, and we urge you to proceed with your plans to encrypt messaging through Facebook products and services. We encourage you to resist calls to create so-called 'backdoors' or 'exceptional access' to the content of users' messages, which will fundamentally weaken encryption and the privacy and security of all users."[46]

As Susan Landau, one of the key authors of both the *Keys Under Doormats* and the *Bugs in Our Pockets* papers, wrote in 2020:

> Law enforcement's line on encryption is that surely the smart people in Silicon Valley can figure out how to build systems that enable law enforcement, backed up with a court order, to access encrypted communications and encrypted data on phones. In reality, such

[43] Bruce Schneier, *Five-Eyes Intelligence Services Choose Surveillance Over Security*, Schneier on Security (Sep. 6, 2018), https://www.schneier.com/blog/archives/2018/09/five-eyes_intel.html.
[44] Stephanie Kirchgaessner et al., *Revealed: Leak Uncovers Global Abuse of Cyber-Surveillance Weapon*, The Guardian (July 18, 2021), https://www.theguardian.com/world/2021/jul/18/revealed-leak-uncovers-global-abuse-of-cyber-surveillance-weapon-nso-group-pegasus.
[45] William J. Lynn III, *Defending a New Domain: The Pentagon's Cyberstrategy*, U.S. Dep't of Defense (2010), https://archive.defense.gov/home/features/2010/0410_cybersec/lynn-article1.aspx (discussing compromises of U.S. classified military networks); *see also, e.g.*, U.S. government Operations "Nitro Zeus" and "Olympic Games" against Iranian nuclear facilities via the Stuxnet malware.
[46] *Open Letter: Facebook's End-to-End Encryption Plans*, Center for Democracy & Technology (Oct. 4, 2019), https://cdt.org/insights/open-letter-facebooks-end-to-end-encryption-plans/.

> surveillance systems are not easy to build—and not easy to build securely. If the CALEA story reveals anything [as discussed in the article], it shows that when companies build in backdoors, hackers, nation-states and criminals will come. That's not the cybersecurity, national security or public safety solution we need.[47]

Further, the UK Information Commissioner's Office (ICO) has also stated its opposition to the introduction of "backdoors":

> Measures that would introduce widespread "backdoors" to encrypted channels or otherwise enable indiscriminate widespread access, would create systemic weaknesses unacceptably undermining security and privacy rights, introducing data protection risks and adding to the overall safety concerns by creating more spaces for harm. We do not support such measures. We welcome the UK Government's support for strong encryption as well as its position that it does not support the development of so-called 'backdoors' in social media platforms to allow access for law enforcement or security agencies.[48]

In sum, there is a broad consensus at an international level that implementing an e2ee system with intentional vulnerabilities or government-mandated scanning tools would be irresponsible, facilitating cybercrime, endangering human rights, and exposing service providers and users alike to material risks to their safety.

Conclusion

The responsibility for safety is a complex question that involves the engagement of the public, businesses, and government alike. Our goal is to prevent as much harm as we possibly can and quickly respond if and when harm does occur. Internet abuse is a constantly evolving landscape

---

[47] Susan Landau, *If We Build It (They Will Break In)*, Lawfare (Feb. 28, 2020), https://www.lawfareblog.com/if-we-build-it-they-will-break.

[48] *A Framework for Analysing End to End Encryption in an Online Safety Context v1 02/11/2021*, UK Information Commissioner's Office (Nov. 2, 2021), https://ico.org.uk/media/about-the-ico/documents/4018823/ico-e2ee-paper-02112021.pdf.

and we are developing the tools to address it. Coordination with governments and stakeholders will be crucial to find solutions and effectively address problems.

Meta is continuing to invest more than any other company in preventing, detecting, and responding to abusive behavior across its products.