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

First published April 2022, and updated September 2023

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] Communications that are e2ee reinforce safety and security and have become the standard expectation of users for their preferred communications platforms. Our plans 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 from being misused to cause harm.

Since 2016, Meta has invested more than $20 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. When they connect with family and friends, 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

[1] Mark Zuckerberg, *A Privacy-Focused Vision for Social Networking*, Meta Newsroom (Mar. 6, 2019), https://about.fb.com/news/2019/03/vision-for-social-networking/.

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 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 privately and securely.

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 and thorough approach to safety. We also need to help protect people from abuse while maintaining the protections that come with encryption. People should be able to protect themselves from unwanted interactions and abuse in a privacy-protected environment. 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 (as described previously) reporting harmful content or behavior from bad actors;
  - Enforcing our Community Standards[2] when we receive reports about a user or content; and
  - 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.

We also believe that metrics and measurements are critical to ensuring the safety, security, and privacy of our messaging services. Meta has developed and refined various metrics to measure our progress in addressing abuse and content that violates our policies, including child sexual exploitation material. For example, we publish metrics about the CyberTip Reports that we make to the NCMEC.[3]

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

---

[2] *Facebook Community Standards*, Meta Transparency Center, https://transparency.fb.com/policies/community-standards/ (last visited Sept. 13, 2023).
[3] *See, e.g.*, *Transparency into Meta's Reports To the National Center for Missing and Exploited Children*, Meta Transparency Center (Sept. 6, 2023), https://transparency.fb.com/en-gb/ncmec-q2-2023/. For example, in Q2 2023, Facebook and Instagram sent over 3.7 million NCMEC CyberTip Reports for child sexual exploitation. Of those reports, 48,000 involved inappropriate interactions with children. *See id.*

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 our platforms, including Messenger and Instagram DMs. Our policies 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 our policies (and the policies 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. Our policies continue to be developed and refined in consultation with global safety experts, including independent online safety organizations and experts.

We know each harm referred to in our policies requires a tailored approach, as explained later in this paper. In developing tailored approaches to the Community Standards[4] harm types, we take a targeted "prevent, control, and respond" approach, whilst also balancing privacy considerations and regulations for Messenger and Instagram DMs.

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:

[4] *See Facebook Community Standards*, Meta Transparency Center, https://transparency.fb.com/policies/community-standards/ (last visited Sept. 14, 2023).

- 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.[5] Research, such as the work of former Federal Bureau of Investigation (FBI) Supervisory Special Agent Ken Lanning for NCMEC in 2010,[6] 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.[7] 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

[5] *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/2021/2/understanding-the-intentions-of-child-sexual-abuse-material-csam-sharers/.
[6] 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.
[7] *See* Andrus *et al.*, *Understanding the intentions of Child Sexual Abuse Material (CSAM) sharers*; Lanning, *Child Molesters: A Behavioral Analysis*.

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. Previous research we conducted on a two-month sampling of our CyberTips found that more than 90% of the reports we make are of content that was the same as or visually similar to previously reported content. And six videos were responsible for more than half of the content reported in those two months. In other words, a small number of images represented the large majority of the images shared (and reshared) and reported. Our research has also found that a very large portion of this content is shared without "malicious intent" — meaning it is shared by people who do not appear to have a sexual interest in children.[8] While the sharing of this content is still harmful, these users are not likely to be the focus of law enforcement investigations.[9]

Preventing Harm at Source

One of the most important steps we can take to keep people safe is by preventing harm from happening in the first place. Our investment in prevention draws on a growing body of research that has recognized the effectiveness of crime and harm prevention.

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 other types of 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 needing to access (unless reported) or scan the contents of our users' private messages, our systems are designed to 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, thereby disrupting potential harm before it happens.

---

[8] *Id.*
[9] Lanning, *Child Molesters: A Behavioral Analysis*.

Our product designs will help divert and deter would-be offenders, limit potentially harmful interactions on our messaging services, and empower users to report suspicious behavior while educating them 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 . 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. The risks posed by diminished security in communications are hardly hypothetical. For example, the UN High Commissioner for Human Rights has reported that at least 65 governments across the globe have acquired commercial spyware surveillance tools, which have been used to target journalists, human rights defenders, politicians, government officials, diplomats, judges, lawyers, doctors, union

leaders, and academics, among others.[10] Such hacking has been linked to the arrest, detention, torture, and even extrajudicial killing of the hacked targets.[11]

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[12] for nefarious ends, as well as threat actors or criminals seeking to gain unauthorized access to the platform's systems.[13]

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:

---

[10] *The right to privacy in the digital age*, A/HRC/51/17, UN High Commissioner for Human Rights (Aug. 4, 2022), https://www.ohchr.org/en/documents/thematic-reports/ahrc5117-right-privacy-digital-age, ¶ 5–6.
[11] *Id.* ¶ 9.
[12] *See, e.g., Former Twitter Employee Found Guilty of Acting as an Agent of a Foreign Government and Unlawfully Sharing Twitter User Information*, U.S. Department of Justice (Aug. 10, 2022), https://www.justice.gov/opa/pr/former-twitter-employee-found-guilty-acting-agent-foreign-government-and-unlawfully-sharing; 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.
[13] *See, e.g.,* Brian Fung, *Twitter Hackers Accessed Direct Messages of 36 Accounts, Company Says,* CNN Business (July 22, 2020), https://www.cnn.com/2020/07/22/tech/twitter-hack-direct-messages/index.html.

> 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.[14]

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. The UN High Commissioner for Human Rights, for example, has written that "[e]ncryption is essential if people are to feel secure in freely exchanging information with others on a range of experiences, thoughts and identities, including sensitive health or financial information, knowledge about gender identities and sexual orientation, artistic expression and information in connection with minority status."[15] Similarly, in a November 2020 statement,[16] 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

---

[14] *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.

[15] *The right to privacy in the digital age*, UN High Commissioner for Human Rights, ¶ 21.

[16] *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/.

plays in free expression, former UN Special Rapporteur for 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."[17] Moreover, because e2ee is so critical to the protection of fundamental rights, efforts to weaken encryption may themselves constitute human rights violations, as some commentators have argued.[18]

For example, while end-to-end encrypted chats are now available as an option on Messenger and Instagram Direct and are by default on WhatsApp, we had previously 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. When appropriate, 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."[19] 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

---

[17] *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, David Kaye*, A/HRC/29/32, UN General Assembly – Human Rights Council (May 22, 2015), https://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/29/32.

[18] *See, e.g.*, Ioannis Kouvakas, *Changes to UK Surveillance Regime May Violate International Law*, Just Security (Aug. 22, 2023), https://www.justsecurity.org/87615/changes-to-uk-surveillance-regime-may-violate-international-law/ (arguing that "the United Kingdom could breach international human rights law by, for example, preventing a communications services provider from … applying advanced protections such as end-to-end encryption to their services, at a global level").

[19] BSR, 2022. "Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption," https://www.bsr.org/reports/bsr-meta-human-rights-impact-assessment-e2ee-report.pdf, at 34.

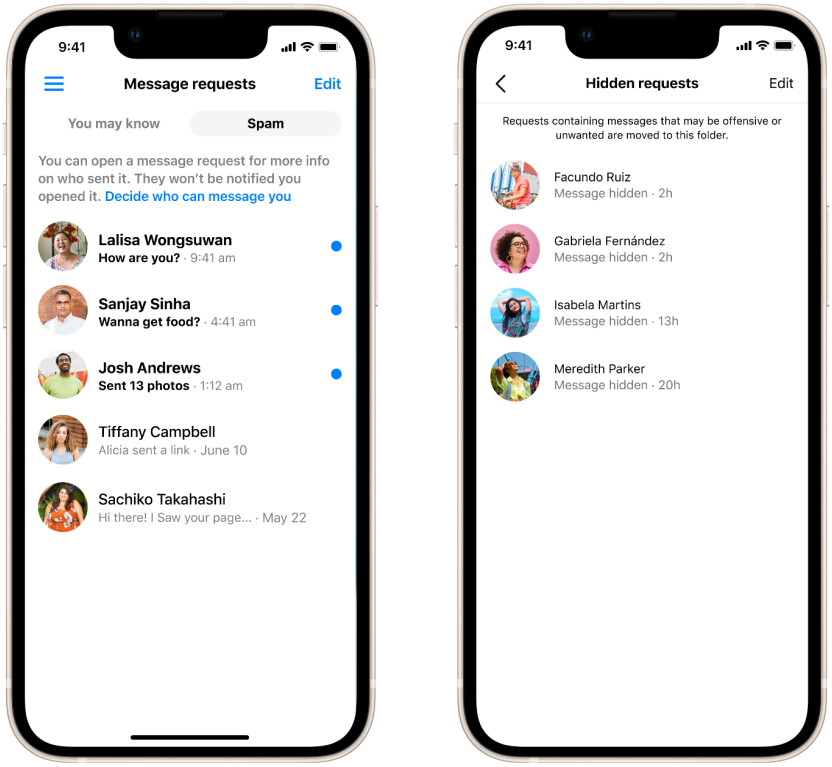
enable these other rights."[20] These groups include investigative journalists, marginalized racial, ethnic, and religious groups, individuals in abusive relationships and victims of trafficking who use messaging platforms to seek help, and civil society organizations, particularly those focused on women and LGBTQI+ rights groups, among others.[21] 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[22] 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."[23]

2. Detecting patterns of potential 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 as far as possible before it reaches messaging. When e2ee is default, we will also use a variety of tools, including artificial intelligence, subject to applicable law,[24] to proactively detect accounts engaged in malicious patterns of behavior instead of scanning private messages.

---

[20] *Id.* at 63.
[21] *Id.*
[22] *Id.* at 33–35.
[23] *Id.* at 44.
[24] Aura Salla, *Changes to Facebook Messaging Services in Europe*, Meta Newsroom (Dec. 20, 2020), https://about.fb.com/news/2020/12/changes-to-facebook-messaging-services-in-europe/.



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.

We hire specialists with backgrounds in law enforcement and work with child safety experts and organizations to identify these actors, monitor their latest tactics and disrupt their networks. Between 2020 and 2022, these teams dismantled 27 abusive networks and in January 2023, we disabled more than 490,000 accounts for violating our child safety policies.

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 is repeatedly blocked or reported by a teen, we can limit that adult from further interactions with teens. Notably, because these behavioral and content signals take place on our public surfaces, we can still use them when our messaging services move to default e2ee.

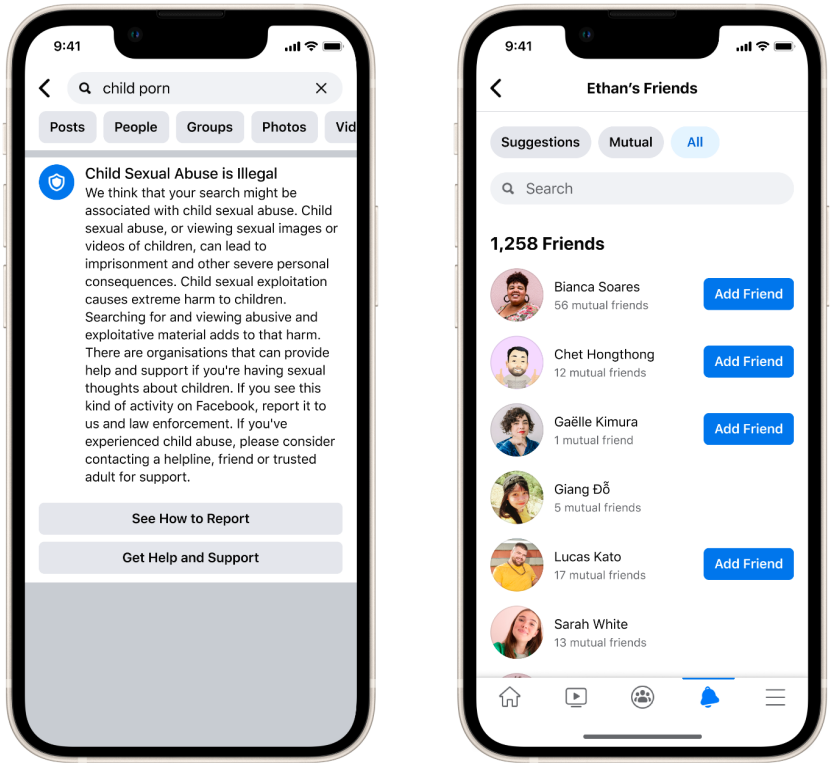
It is important to recognise that bad actors using our platforms to try to find a potential victim or like-minded people to connect with are likely to exhibit suspicious behavior on public surfaces. Signals of harm (which can include blocking, reporting, and the content and metadata we have on our social media platforms, as well as group names, group photos, and metadata we have from Messenger and Instagram DMs, including those where conversations are end-to-end encrypted) can be used as a consistent, encrypted-content agnostic, and long-term basis for applying these protections. Already, we use these public signals as part of risk frameworks that allow us to identify potential malicious actors and proactively take steps to address potential harms.

We look at over 60 different signals to find and label accounts belonging to potentially suspicious adults. In addition to the protections we have in place to restrict these accounts from finding, following or interacting with teens, we are also taking steps to restrict these accounts from interacting with one another and building abusive networks. For example:

- These accounts will not be recommended to each other in Search unless people search the exact username.
- These accounts will not be recommended to each other in 'Accounts You May Follow'.
- If these accounts land on another account belonging to a potentially suspicious adult, they will not be able to follow the account.
- These accounts will not be shown comments by other potentially suspicious adults on public posts, or on their own accounts.

- These accounts will not be shown to one another in surfaces where we recommend content to people, like Explore and Reels.
- Facebook Groups and Pages that have a certain percentage of members or admins that exhibit potentially suspicious behavior will not be suggested to others in places where we make recommendations, for example, Search.

Based on these signals and others, we can take action. We have found through research that when bad actors are simply blocked from connecting with or messaging vulnerable users, they will keep looking for new ways to offend so Meta focuses on limiting bad actors' capabilities within the product experience. This includes prevention actions such as limiting a user's 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 search options on Facebook, preventing a suspicious account from messaging with a minor, or banning accounts entirely).



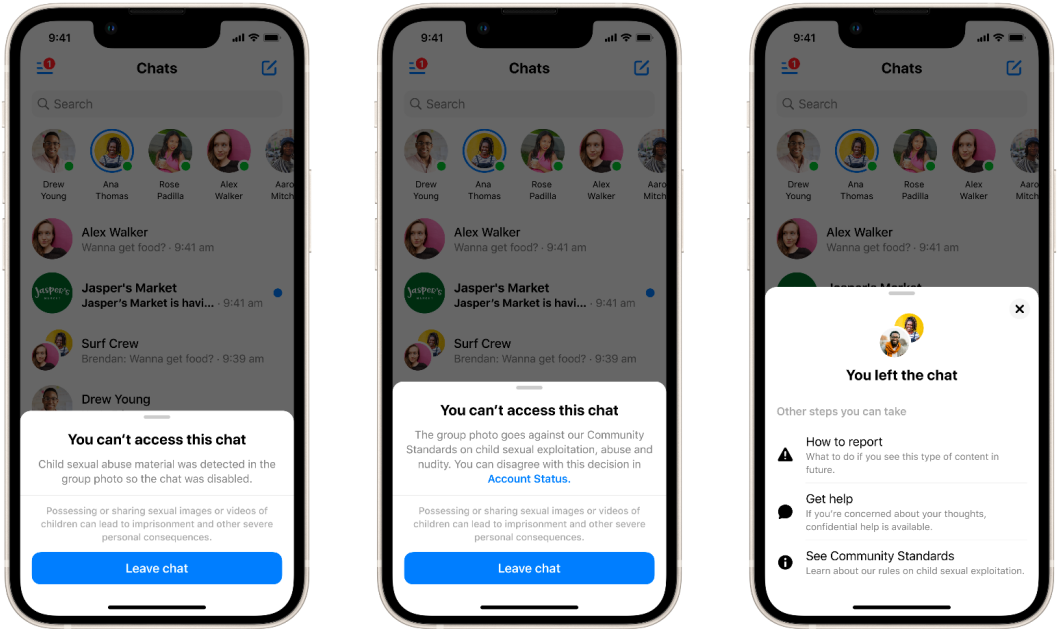
Our approach on redirecting to educational resources is similar to the redirect method used in counter-terrorism.[25] For example, we disrupt and re-direct any user searching for certain CSAM or CSAM-adjacent terms by showing users empty search results as well as showing them information on both 1) how users can report CSAM content to Meta and law enforcement in the case they are a concerned user, and 2) how users can get help with their problematic sexual thoughts and/or behaviors.

We disable accounts that egregiously or repeatedly violate our child safety policies, and we automatically disable Facebook and Instagram accounts if they exhibit a number of signals we monitor for potentially suspicious behavior. We know predators may attempt to set up multiple accounts to evade our enforcement, so when they violate certain child

[25] *The Redirect Method*, Moonshot, https://moonshotteam.com/the-redirect-method/ (last visited Sept. 13, 2023).

safety policies, we disable other accounts held by the account holder, restrict the device from setting up future accounts, and disable any linked Facebook accounts. Between May 27, 2023 and June 2 2023 we automatically blocked more than 29,000 devices on Instagram for child safety violations.

Electronic Service Providers (ESPs) are legally obligated to report apparent violations of laws related to child sexual abuse material (CSAM) they become aware of to NCMEC's CyberTipline. In addition to reporting content we become aware of, we've developed sophisticated technology to proactively seek out this content, and as a result we find and report more CSAM to NCMEC than any other service today. And because many of the signals we use to find and disable accounts belonging to bad actors are e2ee agnostic, we will continue to be able to protect children and make reports to NCMEC in encrypted environments.



Because we recognise this is a constantly evolving area, we work regularly with child safety experts and anti-trafficking organizations to help us understand evolutions in coded language and to identify terms, phrases, slang, and emojis that could be used in an attempt to evade our detection systems and bypass our policies. We use these signals to train our technology so we can proactively find new trends in behavior and take action.[26] Our teams also use technology to find misspellings and spelling variations of this language, as well as terms and phrases related to child exploitation (such as hash tags or links of websites known for sharing) that we can input into our systems to proactively find and disrupt efforts to evade our protections.Additionally, we have hosted or attended regular child safety hackathons since 2016 with our nonprofit partners to further improve our ability to detect potential bad actors. This initiative brings together engineers, data scientists and designers from across the industry who code and prototype projects focused on making the internet a safer place for teens.

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

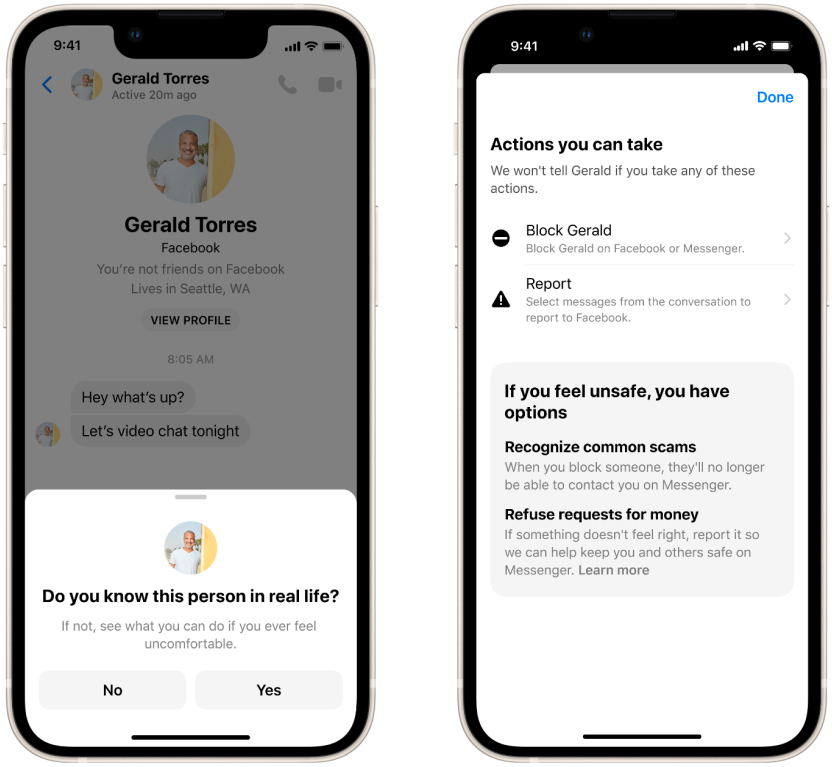
We also aim to prevent harm by empowering and educating users on how to identify and protect themselves from unwanted interactions on Messenger and Instagram DMs through in-app safety notices and easy-to-use reporting tools.

[26] By removing CSAM from Facebook and Instagram, we limit future public or private sharing of this material and prevent benign users from encountering sexualising content. As of Q2 2023, 96.9% of child sexual exploitation content on Facebook was removed proactively by Meta before any user reported it. *See Community Standards Enforcement Report – Child Endangerment: Nudity and Physical Abuse and Sexual Exploitation*, Meta Transparency Center, https://transparency.fb.com/reports/community-standards-enforcement/child-nudity-and-sexual-exploitation/facebook/ (last visited Sept. 15, 2023).

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.[27] 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 just one month in 2023, more than 85 million users saw safety notices on Messenger. And, importantly, both these safety notices and the behavioral signals they rely on are compatible with e2ee.

[27] Jay Sullivan, *Preventing Unwanted Contacts and Scams in Messenger*, Messenger News (May 21, 2020), https://messengernews.fb.com/2020/05/21/preventing-unwanted-contacts-and-scams-in-messenger/.



We want to educate more people to act if they see something and avoid sharing harmful content, even in outrage. Educating users, including teens, not to share CSAM by targeting behaviors associated with viral or sexual content significantly increases reporting in our experiments. 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. Finally, we have worked with public awareness experts to launch the Help Protect Children Campaign, which prevents the sharing of CSAM and encourages users to report such content instead. The campaign educates users on the harm CSAM causes, and the impact it has on victims. It encourages anyone who sees

harmful videos and images of teens to protect the victim by reporting it immediately to Facebook.

Protecting Teens

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. Indeed, Meta has created over 30 tools[28] to support teens on our apps.

For example, we've improved the options for reviewing chat requests and built delivery controls that let users choose who can message their chats list, who goes to their requests folder, and who can't contact them at all. To help users review these chat requests in the safest way possible, we blur images and videos in Messenger and have text only in Instagram[29], we block links, and let users delete requests to chat in bulk. Users can already block unwanted contacts on Instagram and on Instagram Direct, and we now automatically block other accounts the blocked account may have or create (to the best of our ability), while users are given the option to report upon blocking someone. Through this overall approach we have increased reports sent to us by teens in Q1 2023 versus Q1 2022 on Messenger and Instagram DMs by 75%. We are also making it easy to block contacts from strangers — a feature that is switched on by default for any user we identify as a potential teen.

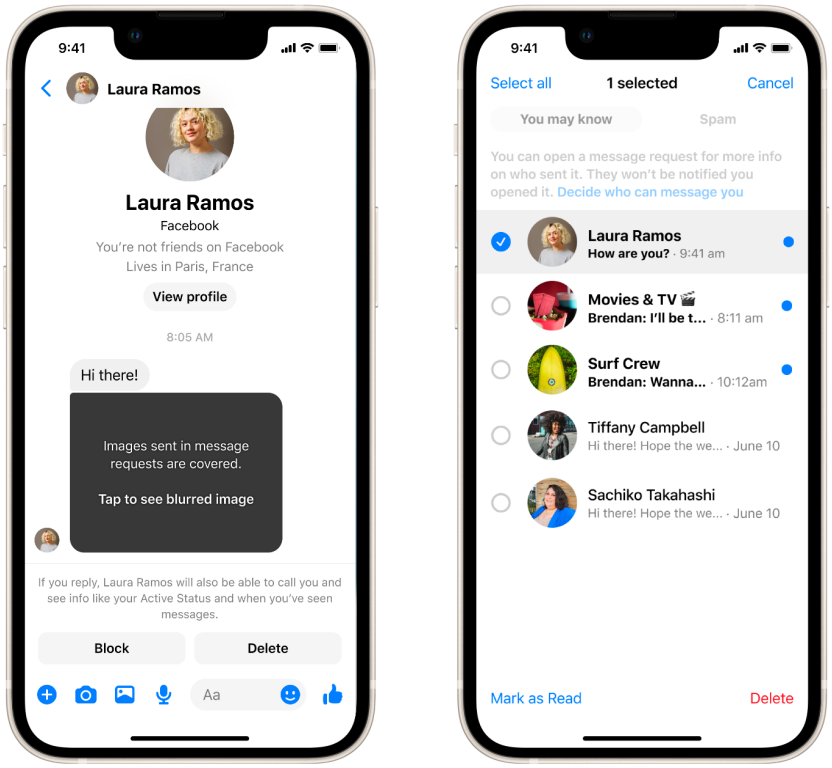
Using machine learning, we're also able to analyze behavioral data across our platforms to proactively identify suspicious signals and take action. We don't allow accounts that exhibit potentially suspicious behavior to find, follow or interact with young people. Specifically, we don't recommend young people's accounts to these individuals, and take further, stricter steps to prevent them from finding and connecting with minors

---

[28] Antigone Davis, *How Meta Is Working to Provide Safe, Age-Appropriate Experiences for Teens*, Meta Newsroom (Jan. 9, 2023), https://about.fb.com/news/2023/01/providing-safe-experiences-for-teens/.

[29] https://about.fb.com/news/2023/06/parental-supervision-and-teen-time-management-on-metas-apps/

altogether or interacting with their content, for example. We see these new efforts are reducing risky conversations between suspicious accounts and minors.



Additional examples of features in place to protect teens include:

- On Facebook and Instagram, setting accounts to private by default for new users under 18 or under 16 (depending on country), meaning that their content cannot be seen by others without permission of the user.
- Providing 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.

- Preventing the profiles of people under 18 years old from appearing in search tools outside of Facebook.
- On Facebook, setting new teen accounts to limit posts to "friends only" by default.
- Removing teens from the "suggested friends" of potentially suspicious adults and removing the ability for certain adults to friend minors.
- Providing users with features like blocking and deleting friend requests.
- Blocking adults over the age of 19 from initiating Instagram DMs with teens, and limiting adults from messaging teens they aren't connected to on Messenger.
- Preventing teens from initiating messaging with unconnected users who we detect are exhibiting risky behaviors on the platform.

- On Messenger, alerting teens of suspicious activity and prompting them to take action to block, report, or ignore/restrict someone when something doesn't seem right through Safety Notices.
- Restricting people under 18 years from accessing certain services supported by Messenger such as Marketplace, Mentorships, Fundraisers, Dating, and Blood Donation.
- Protecting certain information such as contact info, school, or birthday from appearing publicly.
- Making location sharing off by default for all users. When a teen turns on location sharing, we include a consistent indicator as a reminder that they're sharing their location.
- Making it even easier for teens to report content by checking to see if a teen wants to report after blocking or deleting a friend request, making our reporting tools easier to find, and reducing the number of steps to report.
- Creating a new "involves a child" option when reporting certain harms, which helps to prioritize the report for review and action.
- Providing Family center on Messenger[30] which enables teens to get extra support from a parent or guardian when using Messenger. Teens can allow a parent or guardian to supervise their Messenger account, provide extra support and help balance their time. Parents or guardians can: view how much time their teen spends on Messenger; view and receive updates on their teen's Messenger contacts list, as well as their teen's privacy and safety settings; get notified if their teen reports someone (if the teen chooses to share that information); view who can message their teen; and view who can see their teen's Messenger stories[31]. Teens' chats and messaging activity stays private.
- Testing reminding teens of the importance of safe sharing when they open their camera. Early results show fewer teens will share media with risky users.

[30] https://familycenter.meta.com/

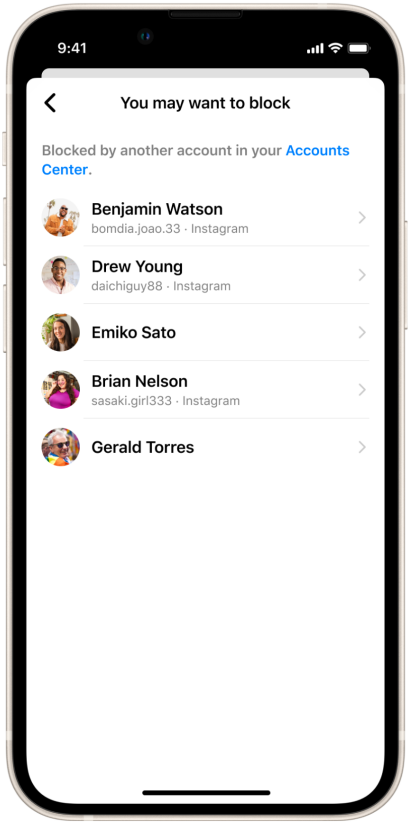
[31] https://about.fb.com/news/2023/06/parental-supervision-and-teen-time-management-on-metas-apps/

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 to teens (described above), we give all users more control of their messaging inbox to account for the variety of experiences people want.

As we've described, 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. As noted above, 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.*)[32]

[32] Salla, *Changes to Facebook Messaging Services in Europe*, 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. For example, on Instagram, we give users the ability to not only block a single account, and we now automatically block other accounts the blocked account may have or create (to the best of our ability), while users are given the option to report upon blocking someone. Based on initial results from this change, we expect that four million fewer accounts will need to be blocked every week since these accounts will now be blocked automatically.

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.[33] The user decides on a list of potentially offensive words, hashtags, and emojis by default, even if they don't break our rules. This will work on e2ee, on-device, under the control of the user. This feature has been effective at keeping people safe. When users with more than 10,000 followers turn on Hidden Words for comments, on average they see 40% fewer comments that might be offensive.

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 law enforcement agencies.

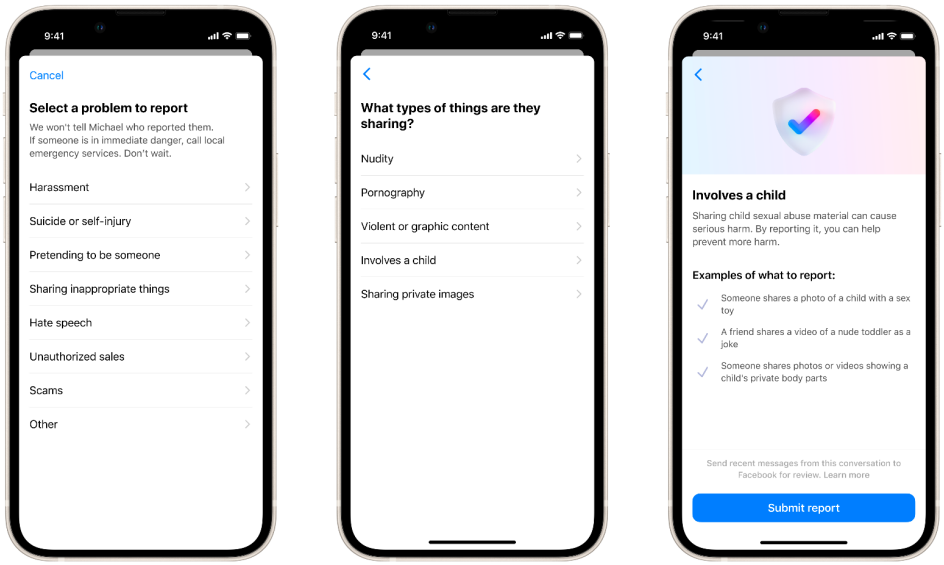
1. User reporting

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, 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.[34] When reporting harm in the vast majority of situations, 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

[33] *Introducing new tools to protect our community from abuse*, Instagram Announcements (Apr. 21, 2021), https://about.instagram.com/blog/announcements/introducing-new-tools-to-protect-our-community-from-abuse.
[34] Antigone Davis, *Preventing Child Exploitation on Our Apps*, Meta Newsroom (Feb. 23, 2021), https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/.

young people. As mentioned above, we have increased reports sent to us by teens in Q1 2023 versus Q1 2022 on Messenger and Instagram DMs by 75% through enhancements such as targeted upsells (report after block) and making reporting flows simpler.



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.[35] 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. These efforts have already yielded

[35] *What end-to-end encryption on Messenger means and how it works*, Facebook Help Center, https://www.facebook.com/help/786613221989782 (last visited Sept. 13, 2023).

positive results, as noted above with regards to increases in teen reporting following the deployment of targeted upsells and simplified reporting flows.

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 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,[36] 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")[37] 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.

---

[36] *Government Requests for User Data*, Meta Transparency Center, https://transparency.fb.com/reports/government-data-requests/ (last visited Sept. 13, 2023).
[37] *Information for Law Enforcement*, Facebook Help Center, https://www.facebook.com/help/494561080557017 (last visited Sept. 13, 2023).; *Law enforcement*, Meta Safety Center, https://about.meta.com/actions/safety/audiences/law (last visited Sept. 13, 2023).

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 81% of law enforcement surveyed cited subscriber information (including name, address, email address, and phone number) and traffic data as the types of data needed most often in investigations.[38] 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.

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,[39]

[38] *SIRIUS EU Digital Evidence Situation Report*, SIRIUS (2022), https://www.europol.europa.eu/cms/sites/default/files/documents/SIRIUS_DESR_2022.pdf, at 27.

[39] 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.

including content brought to our attention by government requests. For example, in Q2 2023, Facebook and Instagram submitted more than 3.7 million NCMEC CyberTip reports for child sexual exploitation.[40] 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.[41]

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.[42] 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 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 is actionable by law enforcement to support people's safety, security, and privacy. And we already know that this reportable data has real world impact. For example, in a report that was recently sent through NCMEC to an EU Member State, we identified from public signals alone that a user was soliciting child abuse images from like-minded people, he

---

[40] *Transparency into Meta's Reports To the National Center for Missing and Exploited Children*, Meta Transparency Center (Sept. 6, 2023), https://transparency.fb.com/en-gb/ncmec-q2-2023/.

[41] *Information for Law Enforcement*, Facebook Help Center, https://www.facebook.com/help/494561080557017 (last visited Sept. 13, 2023).

[42] *How Meta works with law enforcement*, Meta Transparency Center (Jan. 19, 2022), https://transparency.fb.com/policies/improving/working-with-law-enforcement/.

was interested in very young children, and he was in a position of trust, working with refugees. Law enforcement identified the user on the basis of the data we reported, executed a warrant, made an arrest, and went on to find CSAM on devices in his home. And this all occurred without access to the user's inbox.

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 and 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. Though we have been implementing our intent-based technology and have already learned of actionable reports and law enforcement action as a result, 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.

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, 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 Council 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 May 2021, we published an updated 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.

We supported NCMEC in building Take It Down, a global platform for teens who are worried intimate images they created might be shared on public online platforms without their consent. Launched in December 2022, Take It Down helps prevent a teen's intimate images from being posted online and can be used by other companies across the tech industry. We outline more on that below.

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

- **Self-generated teen nudity**

  In recent 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 - Universitatsmedizin 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 2021, we launched a PSA 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.[43]

- **Take It Down Portal**

  Meta has a cross-industry partnership with the aim of combatting the spread of child sexual abuse material. The Take It Down portal, created by NCMEC with Meta's support and launched in December 2022, is a global platform for teens who are worried intimate images they created might be shared on public online platforms without their consent. Take It Down helps prevent a teen's intimate images from being posted online and can be used by other companies across the tech industry. The portal is one step to help remove online nude, partially nude, or sexually explicit photos and videos that were taken before users were 18. Take It Down works by assigning a unique digital fingerprint,

[43] Andrus *et al.*, *Understanding the intentions of Child Sexual Abuse Material (CSAM) sharers*.

> called a hash value, to a video identified by a user as containing nude, partially nude, or sexually explicit images or videos of themself. This all happens without the image or video ever leaving the device; only the hash value will be uploaded to Take It Down and provided to NCMEC. Participating online platforms can use hash values to detect posts of these images or videos on their services, and remove this content. We carry out this process on our non-encrypted surfaces, including Facebook and Instagram, which in turn stops further public and private sharing.

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 to 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."[44]

Given the vitally important role that e2ee plays in protecting individuals' digital privacy, experts have repeatedly highlighted the dangers of undermining e2ee. For instance, in 2022, a group of 70 organizations, cyber security experts, and elected officials signed an open letter to UK Prime Minister Rishi Sunak to warn about the dangers of undermining e2ee:

---

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

> Opening a backdoor for scanning also opens a backdoor for cyber criminals intent on accessing our bank account details, private messages and even the pictures we share online privately with family and friends. We all deserve the protection that end-to-end encryption provides, but the most vulnerable in society – children and members of at-risk communities – need it most of all.[45]

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."[46]

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

---

[45] Ryan Polk, *70 organizations, cyber security experts, and elected officials sign open letter expressing dangers of the UK's Online Safety Bill*, Global Encryption Coalition (Nov. 24, 2022), https://www.globalencryption.org/2022/11/70-organizations-cyber-security-experts-and-elected-officials-sign-open-letter-expressing-dangers-of-the-uks-online-safety-bill/.

[46] 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 (July 6, 2015), http://dspace.mit.edu/bitstream/handle/1721.1/97690/MIT-CSAIL-TR-2015-026.pdf, at 1.

compromise every communication to which the exceptional access mechanism allows access.[47]

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 law enforcement — without the users' consent, control, or knowledge. Following years of debate, including legislative and technical proposals,[48] many of the same security and cryptography experts that had written the "Keys Under Doormats" paper argued in 2021 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."[49]

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

---

[47] *Id.* at 2–3.

[48] *See, e.g.,* Joe Mullin, *It's Back: Senators Want EARN IT Bill to Scan All Online Messages*, Electronic Frontier Foundation (Feb. 3, 2022), https://www.eff.org/deeplinks/2022/02/its-back-senators-want-earn-it-bill-scan-all-online-messages;
Joe Mullin, *The EARN IT Bill Is the Government's Plan to Scan Every Message Online*, Electronic Frontier Foundation (Mar. 12, 2020), https://www.eff.org/deeplinks/2020/03/earn-it-bill-governments-not-so-secret-plan-scan-every-message-online;
A beginner's guide to EU rules on scanning private communications: Part 1, EDRi (Dec. 15, 2021), https://edri.org/our-work/a-beginners-guide-to-eu-rules-on-scanning-private-communications-part-1/;
New Safety Tech Fund Challenge, UK Home Office (Sept. 8, 2021), https://homeofficemedia.blog.gov.uk/2021/09/08/new-safety-tech-fund-challenge/;
Zack Whittaker, *Apple delays plans to roll out CSAM detection in iOS 15 after privacy backlash*, TechCrunch (Sept. 3, 2021), https://techcrunch.com/2021/09/03/apple-csam-detection-delayed/.

[49] Harold Abelson *et al.*, *Bugs in our Pockets: The Risks of Client-Side Scanning* (Oct. 15, 2021), https://arxiv.org/pdf/2110.07450.pdf, at 1.

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 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.

More recently, 68 cyber security academics and researchers set out why surveillance technologies like client-side scanning for removal and reporting would actually undermine safety online as well as privacy for users.[50] Meanwhile, over 300 experts from around the world recently set out at length why scanning technologies and the push for them is "deeply flawed."[51]

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 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."[52] The HRIA emphasizes that security and cryptography experts have raised concerns about the

---

[50] Natasha Lomas, *Security researchers latest to blast UK's Online Safety Bill as encryption risk*, TechCrunch (July 5, 2023), https://techcrunch.com/2023/07/05/uk-online-safety-bill-risks-e2ee/.
[51] *Joint statement of scientists and researchers on EU's proposed Child Sexual Abuse Regulation* (July 4, 2023), https://docs.google.com/document/d/13Aeex72MtFBjKhExRTooVMWN9TC-pbH-5LEaAbMF91Y/edit.
[52] BSR HRIA, at 87.

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."[53] 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.[54]

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.[55] 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 exploited; as Schneier wrote again in 2018, "Demanding that technology companies add backdoors to computers and communications systems puts us all at risk."[56]

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

[53] *Id.* at 86.
[54] *Id.* at 88–89.
[55] Bruce Schneier, *Security vs. Surveillance,* Schneier on Security (Feb. 1, 2016), https://www.schneier.com/blog/archives/2016/02/security_vs_sur.html.
[56] 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.

landscape.[57] History has shown that even the most sensitive and highly secured systems can be breached, even if they are not connected to the internet.[58]

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."[59]

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 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

[57] 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.

[58] William J. Lynn III, *Defending a New Domain: The Pentagon's Cyberstrategy,* U.S. Dep't of Defense (2010), https://apps.dtic.mil/sti/pdfs/ADA527707.pdf (discussing compromises of U.S. classified military networks); *see also, e.g.,* David E. Sanger and Mark Mazzetti, *U.S. Had Cyberattack Plan if Iran Nuclear Dispute Led to Conflict*, The New York Times (Feb. 16, 2016), https://www.nytimes.com/2016/02/17/world/middleeast/us-had-cyberattack-planned-if-iran-nuclear-negotiations-failed.html (discussing U.S. government operations "Nitro Zeus" and "Olympic Games" against Iranian nuclear facilities).

[59] *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/.

> and criminals will come. That's not the cybersecurity, national security or public safety solution we need.[60]

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.[61]

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 and we are developing the tools to address it.

---

[60] 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.

[61] *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.

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.