Meta's response Law enforcement capabilities in relation to child exploitation - Questions on Notice

14 January, 2022

Question 1

CHAIR: You mentioned 600,000. When the AI is identifying these users, is it automatically kicking them off or is it then sending them to a person to look at the account again?

Ms Davis: We review the account, and, if we believe that the account looks to be someone under the age of 13, we then checkpoint them and offer them an opportunity to prove their age before we kick them off.

CHAIR: Are you able to provide us data about how many Australians the AI has identified as being under 13 and who have been removed?

Ms Davis: I don't have that data. I can certainly take back that request.

CHAIR: That would be great, if you are happy to take it on notice; that's quite alright. I understand why you don't have it at your fingertips.

Meta response

Meta takes a multi-layered approach to understanding someone's age - we want to keep people who are too young off of Facebook and Instagram, and make sure that those who *are* old enough receive the appropriate experience for their age.

As per our terms, we require people to be at least 13 years old to sign up for Facebook or Instagram. Understanding people's age on the internet is a complex challenge across industry. Meta has been investing in artificial intelligence tools to help us understand someone's real age, and we've developed technology that allows us to estimate people's ages, like if someone is below or above 18.

We determine a user's age by training our technologies to read multiple signals. In August 2021 we announced that we'll look at things like people wishing friends a happy birthday and the age written in those posts: for example, "Happy 21st Birthday!". We also look at the age users have shared across apps: for example, if a user has shared their birthday on Facebook, we'll use the same for linked accounts on Instagram. We also use this technology to find and remove accounts belonging to people under the age of 13. You can find more information about our approach here https://about.fb.com/news/2021/07/age-verification/

These measures are in addition to the other methods we have in place to find and remove accounts used by people who misrepresent their age. For example, anyone can report an underage account to us. Our content reviewers are also trained to flag reported accounts that appear to be used by people who

are underage. If these people are unable to prove they meet our minimum age requirements, we delete their accounts.

We committed to providing the Committee with updated information on the number of accounts we have removed for not meeting our minimum age requirement. Between July and September 2021, Meta removed more than 2.6 million accounts on Facebook and 850,000 accounts on Instagram globally because they were unable to meet our minimum age requirement.[1] We are working on being able to provide country-specific level figures in the future.

[1] A Mosseri, *Hearing Before the United States Senate Committee on Commerce, Science, and Transportation Subcommittee on Consumer Protection, Product Safety, and Data Security*, 8 December 2021, https://www.commerce.senate.gov/services/files/3FC55DF6-102F-4571-B6B4-01D2D2C6F0D0

Question 2

CHAIR: Can I ask you about your alert services that you're using? This is an idea that's been picked up by a lot of witnesses, and, from the committee's view, it seems to have a lot of merit. When you're sending somebody an alert to indicate that their behaviour might be inappropriate or the conversation that they're having with a child is inappropriate, what we've struggled with from other witnesses is actually quantifying whether there's much data around whether it's working, whereas you guys are using it, and I'm sure you have plenty of research. Does your research show that people are modifying their behaviours as a result and that it does short-circuit some of these conversations and grooming activities, and alerts you to them?

Ms Davis: We've seen from some of the tests that we've been doing—and this is across the range of things that we're doing—an increase in reporting of about 50 per cent year over year. I think there's more measurement and more understanding of what we want to do. I'll give you an example of another tool that we've used that has been quite successful. We have a tool that signals to somebody when they go to post a comment that it may be a comment that is bullying or problematic based on some of the signals we're taking in. When we actually surface up with, 'Hey, this looks like it may not be. This may be bullying. Are you sure you want to actually go forward?' about 50 per cent of the people actually either don't post or modify their comments. I think there are a lot of opportunities to fundamentally change the way people are behaving on our platform.

CHAIR: If you wouldn't mind taking on notice—and the committee has the ability to receive on notice evidence confidentially—any research or figures around behavioural change, particularly around the grooming and CAM aspects of these alert systems, we'd be very grateful to receive it. It would help our thinking about whether this is a tool that could be used more broadly.

Meta response

We have invested considerably to develop tools that keep people safe and secure when they're communicating in private messenger.[2] These tools focus on detecting and deterring potential offenders, and giving users the controls they need to prevent abuse, and educating users on child safety.

Tools to deter potential offenders

Our tools to deter potential offenders are informed by ongoing research. For example, in 2021, we conducted research with global child safety experts - including the National Center for Missing & Exploited Children (NCMEC)- into the intent behind sharing of child sexual abuse material (CSAM).[3] It found that people share these images with a variety of intentions. While our work to understand intent is ongoing, our initial findings tell us that approximately 75 per cent of CSAM sharing on our services is due to people sharing it with non-malicious intent (for example, out of shock or outrage, out of

---

[2] A Davis, 'Our approach to safer private messaging', *Meta Newsroom,* 1 December 2021, https://about.fb.com/news/2021/12/metas-approach-to-safer-private-messaging/

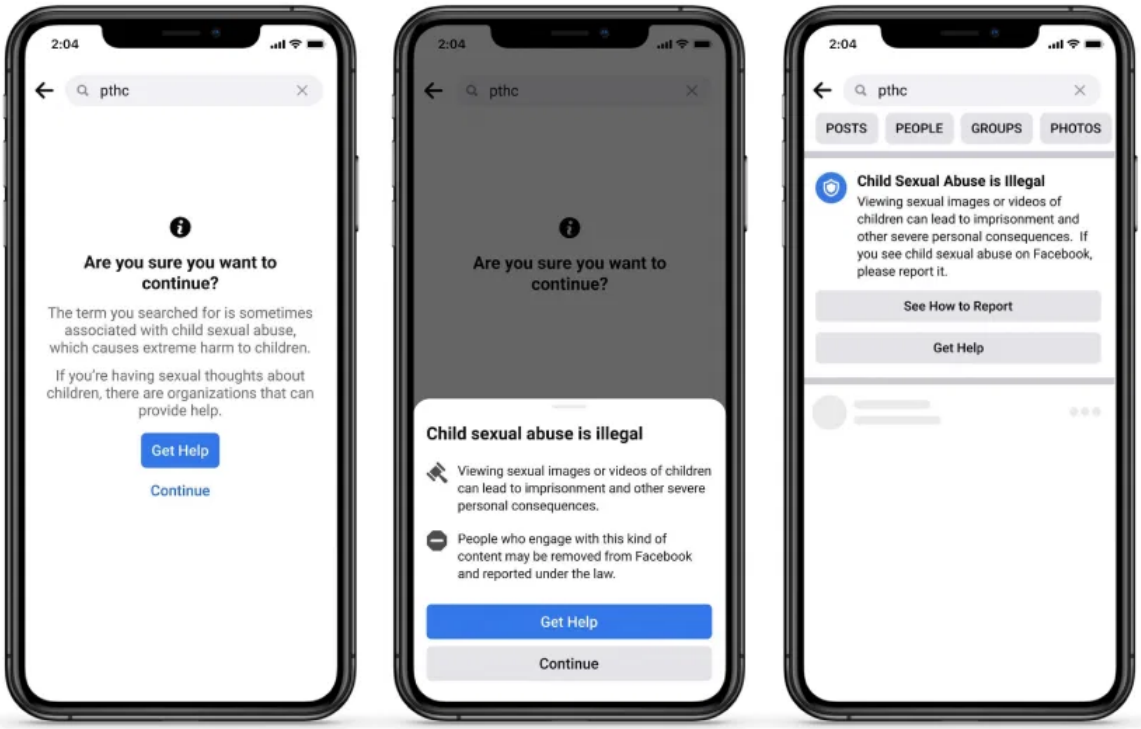
[3] J Buckley, M Andrus and C Williams, 'Understanding the intentions of Child Sexual Abuse Material (CSAM) sharers; *Meta Research Blog*, 23 February 2021, https://research.fb.com/blog/2021/02/understanding-the-intentions-of-child-sexual-abuse-material-csam-sharers/.

ignorance, in poor humour [eg. someone sharing an image of a child's genitals being bitten by an animal], or children sending sexual imagery of themselves to another child).

Based on this research we have introduced two new tools - one aimed at the potentially malicious searching for this content, and another aimed at the non-malicious sharing of this content.[4]

The first is a pop-up that is shown to people who search for terms on our apps associated with child exploitation, shown in Figure 1 below. The pop-up offers ways to get help from offender diversion organisations and shares information about the consequences of viewing illegal content.

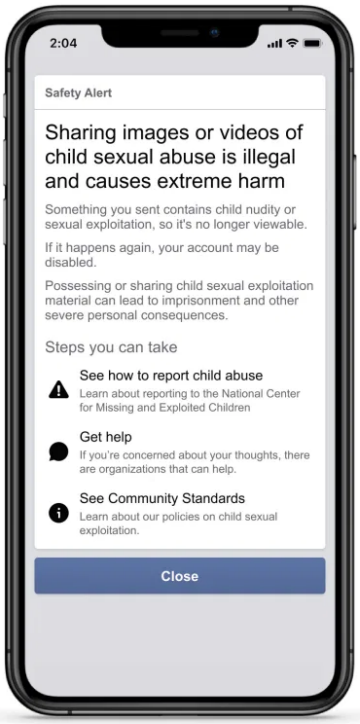
Figure 1: Pop up educating users on the consequences of viewing illegal content



The second is a safety alert, shown in Figure 2 below, that informs people who have shared viral child exploitative content about the harm it can cause, and warns that it is against our policies and there are legal consequences for sharing this material. We share this safety alert in addition to removing the content, banking it and reporting it to NCMEC. Accounts that promote this content will be removed. We are using insights from this safety alert to help us identify behavioral signals of those who might be at risk of sharing this material, so we can also educate them on why it is harmful and encourage them not to share it on any surface — public or private.

[4] A Davis, 'Preventing child exploitation on our apps', *Meta Newsroom*, 23 February 2021, https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/ .

Figure 2: Safety alert warning against sharing illegal material



User controls to prevent potential abuse

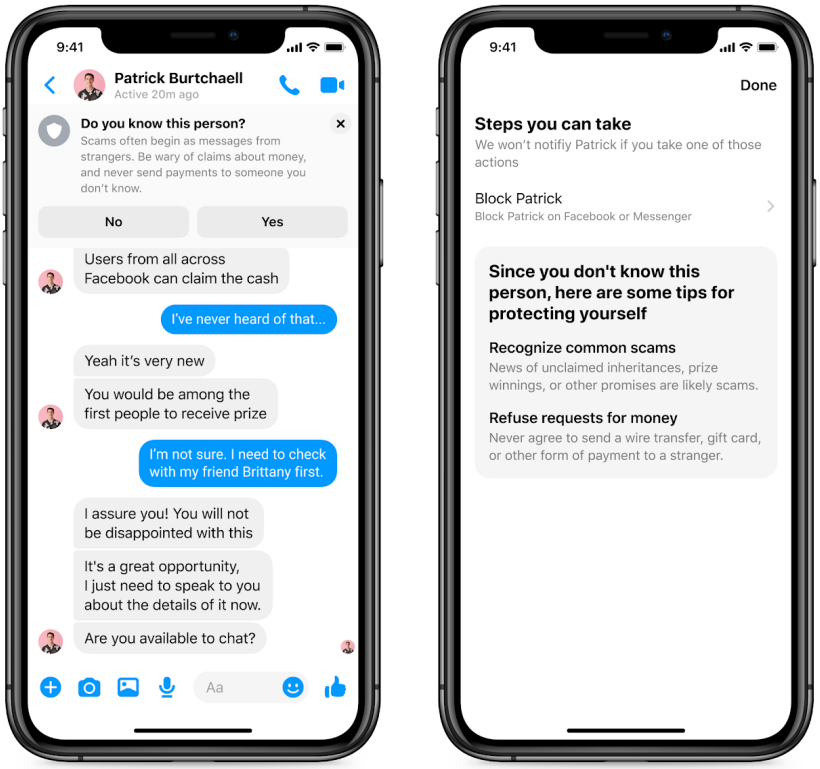
We also educate young people with in-app advice on avoiding unwanted interactions. We've seen tremendous success with our safety notices on Messenger, which are banners that provide tips on spotting suspicious activity and taking action to block, report or ignore/restrict someone when something doesn't seem right.

For example, since 2020, we have sent notices to users in Messenger where we believe an adult could be pursuing a potentially inappropriate private interaction with a child, see Figure 3 below. These are used in instances where someone may be grooming or scamming another user.[5] And, this feature works with end-to-end encryption.[6]

Figure 3: Messenger Safety Notice

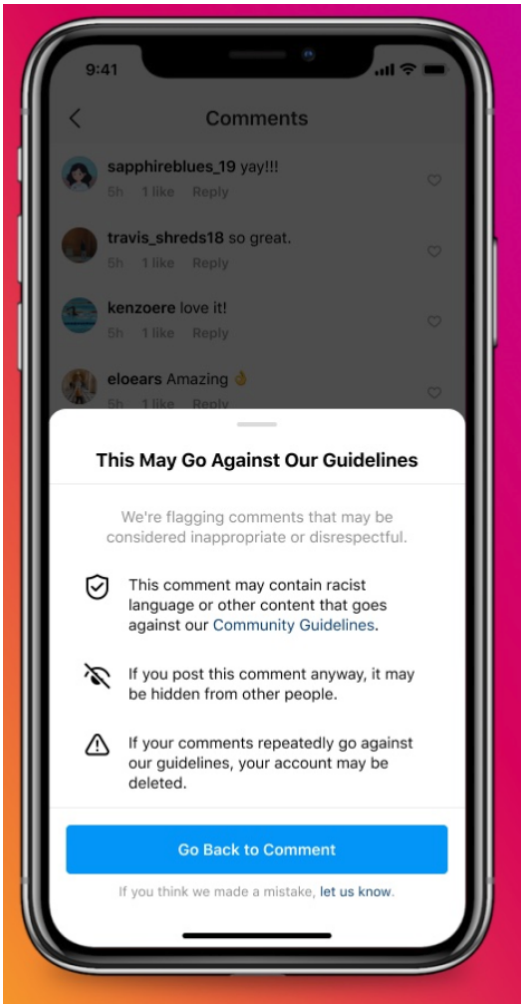
[5] J Sullivan, 'Preventing unwanted contacts and scams in Messenger', *Messenger News*, 21 May 2020, https://messengernews.fb.com/2020/05/21/preventing-unwanted-contacts-and-scams-in-messenger/.

[6] A Davis, 'Our approach to safer private messaging', *Meta Newsroom,* 1 December 2021, https://about.fb.com/news/2021/12/metas-approach-to-safer-private-messaging/



In addition to these tools, we have also recently introduced new interventions to alert users to other potential violations of our Community Standards, such as bullying and harassment. For example, we have recently deployed warning screens on Facebook and Instagram to educate and discourage people from posting or commenting in ways that could be bullying and harassment, shown in Figure 4 below.

Figure 4: Warnings to discourage bullying or harassment



We conduct internal research to measure the success of these interventions, and continue to monitor the role these interventions have on user's behaviour. In the month of November 2021, more than 100 million people have seen safety notice banners on Messenger (Figure 3).[7] Further, we have found that in relation to the bullying and harassment warning screens on Instagram, approximately 50 per cent of the time the comment was edited or deleted by the user based on these warnings (Figure 4).[8]

Educating users on child safety

[7] A Davis, 'Our approach to safer private messaging', *Meta Newsroom,* 1 December 2021, https://about.fb.com/news/2021/12/metas-approach-to-safer-private-messaging/

[8] A Davis, 'Our approach to addressing bullying and harassment', *Meta Newsroom,* 9 November 2021, https://about.fb.com/news/2021/11/how-meta-addresses-bullying-harassment/

In addition to surfacing resources and safety information in our products, we also recognise that reducing sharing of CSAM requires a wider, societal level efforts. Based on this, we are developing campaigns to proactively promote child safety.
In light of our findings that the vast majority of people do not share this content with malicious intent, we launched a public safety campaign (PSA) called "Report it, Don't Share it" to remind people not to reshare CSAM online, even in the context of outrage or condemnation, as it causes further harm to the child.

We've launched the campaign in more than 15 countries so far and the video is available in 12 languages. The PSA has been seen over 50 million times worldwide.

We worked with partners across the globe to amplify the 'Report it, Don't Share it' message. In Australia, we launched the campaign during 2021 National Child Protection Week with support from the NCMEC, the Australian eSafety Commissioner and NSW Police. We also held a virtual panel with the Carly Ryan Foundation, NSW Police and ReachOut to discuss tools, tips and resources for parents to keep their children safe online.[9]

[9] M Garlick, 'Every child, in every community, needs a fair go', *Facebook Australia Blog*, 7 September 2021, https://australia.fb.com/post/every-child-in-every-community-needs-a-fair-go/; Meta, 'National Child Protection Week panel', *Facebook*, 7 September 2021, https://www.facebook.com/watch/?v=147823834189789