Instagram to scan under-18s' messages to protect against 'inappropriate images'

Publication Date: 2024-01-25

Author: Alex Hern

Section: Technology

Tags: Instagram, Meta, Facebook, Social networking, Internet safety, Privacy, news

Article URL: https://www.theguardian.com/technology/2024/jan/25/instagram-to-scan-under-19s-messages-to-protect-against-inappropriate-images

Instagram will begin scanning messages sent to and from under-18s to protect them from "inappropriate images", Meta has announced. The feature, being kept under wraps until later this year, would work even on encrypted messages, a spokesperson said, suggesting the company intends to implement a so-called client-side scanning service for the first time. But the update will not meet controversial demands for inappropriate messages to be reported back to Instagram servers. Instead, only a user's personal device will ever know whether or not a message has been filtered out, leading to criticism of the promise as another example of the company "grading its own homework". "We're planning to launch a new feature designed to help protect teens from seeing unwanted and potentially inappropriate images in their messages from people they're already connected to," the company said in a blogpost, "and to discourage them from sending these types of images themselves. We'll have more to share on this feature, which will also work in encrypted chats, later this year." It is the latest in a series of changes Meta has proposed to respond to criticism that plans to encrypt direct messages on Facebook Messenger and Instagram could place children and young people at risk. The broad description of the intended feature is similar to a setting called "communication safety" introduced by Apple in 2023, which detects nude photos and videos sent to children's devices and automatically blurs them, but offers the child the option of viewing them or contacting a trusted adult. The plans stop short of the stronger versions of client-side scanning that children's safety groups have called for, which would more aggressively report such inappropriate messages to the service's moderators, enabling repeat offenders to be tracked and caught. The move came as Meta faces a lawsuit in New Mexico accusing it of failing to protect children on its platforms. According to an unsealed legal filing related to the case last week, Meta estimates about 100,000 children using Facebook and Instagram receive online sexual harassment each day. On Wednesday, Mark Zuckerberg, Meta's chief executive, will appear in front of the US Congress with a number of other social media bosses for a hearing about child safety. Alongside the promise of future scanning tools, Instagram announced a small set of immediate updates to teenager safety features on the platform. Under-19s will now default to privacy settings that prevent anyone they do not follow from sending them direct messages. Previously, the restriction had only applied to adults messaging teens. Other new features are launching for parents using the service's "supervision" tools, which lets them connect their Instagram accounts to their children's to set time limits, monitor their teens' blocks, and be notified when their settings are changed. Now, parents will be prompted to actively approve or deny attempts by children under 16 to loosen safety settings. "As with all our parental supervision tools, this new feature is intended to help facilitate offline conversations between parents and their teens, as they

navigate their online lives together and decide what's best for them and their family," Meta said. Arturo Béjar, a former senior engineer and consultant at Meta, said the changes needed to be accompanied by regular updates on how many unwanted advances teenagers had received on Instagram. Without such data there would be no way of gauging the impact of safety updates. According to Béjar's own research on Instagram users in 2021, one in eight children aged 13-15 on Instagram had received unwanted sexual advances. "This is another 'we will grade our own homework' promise," he said. "Until they start quarterly reporting on unwanted advances, as experienced by teens, how are we to know they kept their promise or the impact it had? Today, after over two years of Meta knowing that each week one in eight kids get unwanted advances, there is still no way for a teen to flag or report an unwanted advance."