Teens on social media need both protection and privacy. AI could help get the balance right

January 31 2024, by Afsaneh Razi

Credit: Unsplash/CC0 Public Domain

Meta announced on Jan. 9, 2024, that it will protect teen users by blocking them from viewing content on Instagram and Facebook that the company deems to be harmful, including content related to suicide and eating disorders. The move comes as federal and state governments have increased pressure on social media companies to provide safety measures for teens.

At the same time, teens turn to their peers on social media for support that they can't get elsewhere. Efforts to protect teens could inadvertently make it harder for them to also get help.

Congress has held numerous hearings in recent years about social media and the risks to young people. The CEOs of Meta, X—formerly known as Twitter—TikTok, Snap and Discord are scheduled to testify before the Senate Judiciary Committee on Jan. 31, 2024, about their efforts to protect minors from sexual exploitation.

The tech companies "finally are being forced to acknowledge their failures when it comes to protecting kids," according to a statement in advance of the hearing from the committee's chair and ranking member, Senators Dick Durbin (D-Ill.) and Lindsey Graham (R-S.C.), respectively.

I'm a researcher who studies online safety. My colleagues and I have been studying teen social media interactions and the effectiveness of platforms' efforts to protect users. Research shows that while teens do face danger on social media, they also find peer support, particularly via direct messaging. We have identified a set of steps that social media platforms could take to protect users while also protecting their privacy and autonomy online.

What kids are facing

The prevalence of risks for teens on social media is well established. These risks range from harassment and bullying to poor mental health and sexual exploitation. Investigations have shown that companies such as Meta have known that their platforms exacerbate mental health issues, helping make youth mental health one of the U.S. Surgeon General's priorities.

Much of adolescent online safety research is from self-reported data such as surveys. There's a need for more investigation of young people's real-world private interactions and their perspectives on online risks. To address this need, my colleagues and I collected a large dataset of young people's Instagram activity, including more than 7 million direct messages. We asked young people to annotate their own conversations and identify the messages that made them feel uncomfortable or unsafe.

Using this dataset, we found that direct interactions can be crucial for young people seeking support on issues ranging from daily life to mental health concerns. Our finding suggests that these channels were used by young people to discuss their public interactions in more depth. Based on mutual trust in the settings, teens felt safe asking for help.

Research suggests that privacy of online discourse plays an important role in the online safety of young people, and at the same time a considerable amount of harmful interactions on these platforms comes in the form of private messages. Unsafe messages flagged by users in our dataset included harassment, sexual messages, sexual solicitation, nudity, pornography, hate speech and sale or promotion of illegal activities.

However, it has become more difficult for platforms to use automated technology to detect and prevent online risks for teens because the platforms have been pressured to protect user privacy. For example, Meta has implemented end-to-end encryption for all messages on its platforms to ensure message content is secure and only accessible by participants in conversations.

Also, the steps Meta has taken to block suicide and eating disorder content keep that content from public posts and search even if a teen's friend has posted it. This means that the teen who shared that content would be left alone without their friends' and peers' support. In addition, Meta's content strategy doesn't address the unsafe interactions in private conversations teens have online.

Striking a balance

The challenge, then, is to protect younger users without invading their privacy. To that end, we conducted a study to find out how we can use the minimum data to detect unsafe messages. We wanted to understand how various features or metadata of risky conversations such as length of the conversation, average response time and the relationships of the participants in the conversation can contribute to machine learning programs detecting these risks. For example, previous research has shown that risky conversations tend to be short and one-sided, as when strangers make unwanted advances.

We found that our machine learning program was able to identify unsafe conversations 87% of the time using only metadata for the conversations. However, analyzing the text, images and videos of the conversations is the most effective approach to identify the type and severity of the risk.

These results highlight the significance of metadata for distinguishing unsafe conversations and could be used as a guideline for platforms to design artificial intelligence risk identification. The platforms could use high-level features such as metadata to block harmful content without scanning that content and thereby violating users' privacy. For example, a persistent harasser who a young person wants to avoid would produce metadata—repeated, short, one-sided communications between unconnected users—that an AI system could use to block the harasser.

Ideally, young people and their care givers would be given the option by design to be able to turn on encryption, risk detection or both so they can decide on trade-offs between privacy and safety for themselves.

This article is republished from The Conversation under a Creative Commons license. Read the original article.

Provided by The Conversation

Citation: Teens on social media need both protection and privacy. AI could help get the balance right (2024, January 31) retrieved 29 April 2024 from https://techxplore.com/news/2024-01-teens-social-media-privacy-ai.html