The biggest AI companies agree to crack down on child abuse images

Tech companies like Google, Meta, OpenAI, Microsoft, and Amazon committed today to reviewing their AI training data for child sexual abuse material (CSAM) and removing it from use in any future models.

The companies signed on to a new set of principles meant to limit the proliferation of CSAM. They promise to ensure training datasets do not contain CSAM, to avoid datasets with a high risk of including CSAM, and to remove CSAM imagery or links to CSAM from data sources. The companies also commit to "stress-testing" AI models to ensure they don't generate any CSAM imagery and to only release models if these have been evaluated for child safety.

Other signatories include Anthropic, Civitai, Metaphysic, Mistral AI, and Stability AI.

Generative AI has contributed to increasing concerns over deepfaked images, including the proliferation of fake CSAM photos online. Stanford researchers released a report in December that found a popular dataset used to train some AI models contained links to CSAM imagery. Researchers also found that a tip line run by the National Center for Missing and Exploited Children (NCMEC), already struggling to handle the volume of reported CSAM content, is quickly being overwhelmed by AI-generated CSAM images.

The anti-child abuse nonprofit Thorn, which helped create the principles with All Tech Is Human, says AI image generation

can impede efforts to identify victims, create more demand for CSAM, allow for new ways to victimize and re-victimize children, and make it easier to find information on how to share problematic material.

In a blog post, Google says that in addition to committing to the principles, it also increased ad grants for NCMEC to promote its initiatives. Google’s vice president of trust and safety solutions, Susan Jasper, said in the post that supporting these campaigns raises public awareness and gives people tools to identify and report abuse.

https://www.theverge.com/2024/4/23/24138356/ai-companies-csam-thorn-training-data