Google

Published April 2025

Progress update: Responsible AI and Child Sexual Abuse and Exploitation Online

Introduction

Working together to combat AI-facilitated child sexual abuse and exploitation.

Google is committed to a bold, responsible, and collaborative approach to AI that maximizes the technology's benefits while minimizing potential harm. A key priority is protecting against, and responding to, new and unique risks for potential child sexual abuse and exploitation (CSAE) that generative AI might pose. We're committed to tackling CSAE, and we prohibit storing, sharing, or creating child sexual abuse material (CSAM) on our platforms and services. This paper outlines the child safety protections Google utilizes in its Responsible AI approach, laid out in the format used in the Safety by Design for Generative AI: Preventing Child Sexual Abuse principles developed by Thorn and All Tech Is Human.

We invest heavily in fighting CSAE online and employ a combination of automated detection tools and specially trained reviewers working around the clock to deter, detect, remove, and report content that is illegal or violates our policies on our platform. This includes technology-facilitated CSAE.

Using a combination of hash-matching technology, classifiers, and human reviews, we act on several kinds of CSAE content and apply our decades of experience responding to the misuse of emerging technologies in the abuse of children online. Our approach includes identifying and appropriately reporting obscene visual representations (OVR) of children, such as CSAM cartoons, and computer-generated imagery (CGI) CSAM. These types of abuse material are created using technology such as image editing and the AI generation of CSAM. In 2024, we reported hundreds of thousands of pieces of CGI or OVR CSAM that appeared on our platforms to the National Center for Missing & Exploited Children (NCMEC), a US non-profit that works with law enforcement and child-serving global organizations to battle online suspected child sexual exploitation, including AI-generated CSAM (AIG) CSAM.

Introduction (cont.)

As we develop helpful AI products, including those that offer creative and educational benefits, we work proactively to identify and respond to CSAE risks. Our commitment to child safety extends to how we research, develop, and launch our newest generative AI tools. We work with non-profits such as the Internet Watch Foundation (IWF) to understand how CSAE harms are emerging through new technologies, and partner with organizations such as Thorn to establish and develop best practices for mitigating CSAE risks and to support the development of safety standards. Our approach, alongside our partners, focuses on emerging AI-facilitated CSAE risks that adversarial actors could create at scale, such as AIG CSAM that violates our prohibited use policy by manipulating existing images and videos into new abusive content, transforming innocent material into sexualized images, or creating entirely AI-generated material.

Our efforts to combat AI-facilitated CSAE are part of our broader end-to-end commitment to develop and deploy generative AI responsibly. We integrate child safety mitigations and considerations at every stage of the AI development lifecycle. This approach focuses on identifying and mitigating potential risks and harms to foster a responsible approach to AI development. We believe that this ongoing process is essential for creating a digital environment that is both safe and beneficial for children. Our work on child safety is constantly evolving to address the risks in this space, including collaborating with the industry to tackle this problem at scale, in order to make the internet a safer place for everyone.

We joined other technology companies in committing to the Safety by Design for Generative AI: Preventing Child Sexual Abuse principles developed by Thorn and All Tech Is Human. These focus on embedding protections against AI-facilitated CSAE from the start and throughout the AI lifecycle in three key phases: Develop, Deploy, and Maintain. Here are some examples of how we are applying these principles:

Develop

Develop, build, and train generative AI models that proactively address child safety risks.

Building for safety means integrating child safety considerations into the development of a model. We've developed a rigorous process for CSAE filtering at multiple stages during data preparation to exclude content that violates our policies for child safety. For example, in developing our Gemma open models, we applied key data cleaning and filtering methods to the training data, such as CSAM filtering, at multiple stages in the data preparation process to exclude harmful and illegal content.

Deploying content provenance tools across our products: In 2024, Google joined the Coalition for Content Provenance and Authenticity (C2PA) as a steering committee member, where we are partnering with others in the industry to develop interoperable provenance standards and technology to explain how content was produced. Once models are deployed into products and services, applications that generate audiovisual content are required to incorporate a robust provenance solution like SynthID. These requirements are based on the nature of the product, its intended user base, planned capabilities, and the types of output involved. For example, an application made available to minors may have additional requirements in areas such as parental supervision and age-appropriate content.

Training data: Many of our training datasets are taken from sources already built or integrated with child safety protections, such as Google Search, which reported and removed over 1 million URLs from the Search index for CSAM in 2024 — over 400,000 URLs between January and June 2024 and over 880,000 between July and December 2024.

In addition, we focus on safeguarding our training datasets from CSAE by integrating hash-matching and child safety classifiers to remove CSAM as well as other exploitative and illegal content from our training datasets. Already in 2025, Google has detected and removed a number of instances of CSAM in training datasets during its preparation of these datasets pre-model training, which were subsequently reported to NCMEC.

Leveraging our Trust and Safety expertise: Google leverages and integrates the expertise developed by our Trust and Safety teams on content safety, privacy, and security into our technology. This can include our pre-launch risk assessment, which identifies which generative AI applications have sufficiently great or novel child safety risks that require specialized testing and controls. It also includes employing guardrails that can include safety filters and tuning, as well as system instructions, in our most capable foundation and frontier models and generative AI applications. Safety filters can act as a barrier, preventing harmful outputs that do not directly influence a model's behavior. Non-configurable safety filters, used to block CSAM and other information such as personally identifiable information (PII), automatically block certain types of harmful content and do not allow users to adjust settings or thresholds for what is

Develop (cont.)

considered unsafe content. This reduces the risk of generating harmful content per our policies for combating CSAE and CSAM.

Conducting multi-layered red teaming and child safety adversarial testing: Before releasing any models publicly, we conduct thorough testing to the extent legally permissible to identify and resolve potential vulnerabilities, and work to mitigate the possibility of CSAE material being produced or our products being misused to enable CSAE. We conduct adversarial child safety testing across text, image, video, and audio for potential risks and violations.

We have developed an evaluation approach that includes red teaming, to the extent legally permissible, and child safety adversarial testing. By inserting adversarial prompts into a model or product and evaluating outputs, we can provide data and feedback to developers. To conduct these child safety evaluations, we developed over 19,000 adversarial prompts using resources, such as intel reports, synthetic prompt development, and subject-matter expertise, to target known risk vectors. Prompts fit a range of modalities, such as text prompts focusing on grooming or image generation focused on sexualized images, and are developed based on the capabilities of the model or product. In 2024, these exercises resulted in the evaluation of over 700,000 model responses.

These techniques play a critical role in our approach to proactively testing AI systems for weaknesses and identifying emerging risks. Upon public release of our models, Google continues to monitor adversarial trends and test our models for additional risks and adversarial pivots that may emerge; for example, by using intelligence vendors or social monitoring. Our approach is continually evolving, incorporating new measurement techniques as they become available as well as insights from resources such as intel reports.

Deploy

Release and distribute generative AI models after they have been trained and evaluated for child safety, providing protections throughout the process.

Protecting against CSAE risks continues beyond the development of a model and into how it is responsibly hosted.

Detecting and responding to CSAE: We incorporate several preventative measures to detect and respond to CSAE content, including using machine learning to identify CSAE-seeking prompts and uploads, blocking models from producing exploitative outputs, and reporting any confirmed CSAM matches to NCMEC. For example, in 2024, we reported more than 600 instances of apparent CSAM to NCMEC using hash-matching, which were uploaded as a user prompt to our generative AI products.

Feedback loops: Alongside our work on filtering, our feedback loops help identify safety issues that can be improved. Users have the option of using these mechanisms, such as Gemini's in-product report a problem feature or Google's report content feature, to report any issues.

Setting clear policies for responsible AI use: Our approach to AI is grounded in our AI Principles, which guide the safety and reliability of Google AI products, focusing on oversight, due diligence, and feedback mechanisms. These principles ensure we align with user goals, social responsibility, and widely accepted principles of international law and human rights.

Our policies and procedures for mitigating harm in areas such as child safety have been informed by years of research, user feedback, and expert consultation. These policies guide our models and products to minimize certain types of harmful outputs and dictate behavior that is prohibited on our products.

As part of our responsible AI approach, we have developed policies that prohibit CSAE on Google's generative AI products, iterating these policies as both the technology and the risk landscape evolve. Our prohibited use policy states: "Do not engage in dangerous or illegal activities, or otherwise violate applicable law or regulations. This includes generating or distributing content that relates to child sexual abuse or exploitation."

Deploy (cont.)

Google products that use generative AI or may process generative AI content also prohibit the use of their technology to abuse or exploit children via product-specific policies. For example, our policy guidelines for the Gemini app include threats to child safety, stating that Gemini should not "generate outputs, including Child Sexual Abuse Material, that exploit or sexualize children." Google Play's Play Console Help Center contains a section focused on understanding Google Play's AI-Generated Content policy that outlines violative AI-generated content, such as AI-generated non-consensual deepfake sexual material, and prohibits content that may exploit or abuse children. Google Cloud Platform's Acceptable Use Policy requires users to agree not "to engage in, promote or encourage illegal activity, including child sexual exploitation, child abuse, or terrorism or violence that can cause death, serious harm, or injury to individuals or groups of individuals."

Maintain

Maintain model and platform safety by continuing to actively understand and respond to child safety risks.

Google works closely with child safety experts, NGOs, industry partners, and law enforcement to combat CSAE. These collaborations take various forms, such as at our annual Google Safety Engineering Center Growing up in the Digital Age summit that brings together 200+ experts, and often result in an increased understanding of the child safety landscape, identification of emerging trends and risks associated with new technology, recognizing developing patterns of abuse, and innovative solutions.

Identifying current, emerging, and potential future AI risks

We develop rigorous research with partners such as the Digital Trust & Safety Partnership (DTSP) and support research efforts spearheaded by organizations such as the Tech Coalition, as well as creating our own intelligence reports to further investigate emerging risks.

Using research-backed insights to enhance our approach: Between 2023 and 2024, Google commissioned a number of internal reports focused on CSAE with an external provider of expert intelligence insights, with several focused specifically on generative AI and how bad actors are using or planning on using AI in abusing children. These reports were then used to inform our approach in a variety of ways, including to update adversarial prompts for child safety evaluations and in further understanding and addressing the risks identified.

Researching best practices and new solutions, including using AI, for addressing CSAE: AI offers significant potential to enhance the moderation of harmful CSAE content and improve child safety. While human oversight remains crucial, research suggests that AI can improve efficiency in tackling CSAE while reducing human reviewers' exposure to content that could cause psychological harm. Such an approach can allow human reviewers to focus on the more nuanced cases that require manual review, with more tools to support their work. We work with the DTSP to research and develop industry best practices for using AI itself towards detecting and removing policy-violative generative AI content, such as AIG CSAM.

Working with industry and government to create and share child safety tools

Collaborating with partners and industry as part of the Tech Coalition: During 2023 and 2024, as a member of the Tech Coalition — a global alliance of technology companies working to end online CSAE — we led several working groups focused on understanding child safety risks in generative AI. This culminated in US, UK, and EU briefings designed to bring together key stakeholders and identify opportunities when it comes to combating AI-facilitated CSAE. As a result of the collaboration in these

Maintain (cont.)

working groups, the Tech Coalition developed member resources including a reporting template for industry reports of AI-generated CSAE to NCMEC, now available to over 45 Tech Coalition member companies. In addition, through its membership, Google also supports the Tech Coalition's research initiatives, including those focused on understanding AI and CSAE risks. For example, the Tech Coalition announced it will fund new research through its Tech Coalition Safe Online Research Fund focused on generative AI and CSAE.

Robust Open Online Safety Tools (ROOST): ROOST is a new cross-industry initiative launched at the Paris AI Action Summit, which aims to build scalable, interoperable safety infrastructure ready for the AI era. We are collaborating with companies such as OpenAI, Discord, and Roblox through the ROOST initiative to share technology and help organizations of all sizes create safer online platforms. This includes providing access to AI safety tools for detecting, reviewing, and reporting CSAE that will work to close the digital safety gap.

Google's Child Safety Toolkit: Since 2014, Google partners have used our free Child Safety Toolkit to analyze billions of images and videos each month for potential CSAE. The toolkit consists of the Content Safety API and CSAI Match, which help with better prioritization, quicker identification, and safer operations. The toolkit is used by companies including Snap and Adobe. In October 2024, the Content Safety API was updated to use semantic data from images, which offers a way for external partners that either cannot send raw image bytes to Google, or have high-volume requirements to leverage our CSAM detection tooling.

Google Priority Flagger Program: As part of our Priority Flagger Program, we partner with expert third parties that flag potentially violative content, including content that raises child safety issues, for our teams' review.

Complying with regulatory reporting requirements in order to identify and assess risks: We are complying with regulatory requirements, such as the European Union's Digital Services Act (DSA) under which we conduct Systemic Risk Assessments (SRAs) to help make the internet more safe, transparent, and accountable. These reports identify and address risks, including CSAE on online platforms. The recently released 2024 SRA notes that generative AI may be used by bad actors to create new outputs that exacerbate some existing child safety risks and introduce new risks. Mitigations for these risks include our product policies, testing our generative AI products before launch, our Child Safety Toolkit, and ongoing proactive engagement with child safety experts from industry, academia, government, and civil society, including our annual Growing Up in the Digital Age summit.

Working with industry experts to quickly spot and remove CSAE

We partner with the Internet Watch Foundation (IWF), an independent non-profit focused on removing online child sexual abuse imagery. The IWF sends Google rapid alerts when criminal content is identified on its platforms, making swift action possible.

"Since the IWF first started monitoring AI-generated child sexual abuse and exploitation (AI CSAE) in early 2023 we've seen a rapid improvement in the ability to generate lifelike imagery.

In 2024, almost all the reports or URLs that IWF analysts dealt with that contained AI-generated CSAE were found on publicly available spaces on the clear web.

This is why we're grateful to technology partners such as Google that are committed to tackling AI-facilitated CSAE. We have worked closely to establish processes that allow Google to receive near-instant alerts if criminal content is flagged on its services, including child sexual abuse material that has been generated by AI.

Generative AI is having a real-world effect now on victims and we need responsive partners to help stave off the devastating impact that misuse of generative AI is having on global child protection. By working together with Google we can aim to have a future internet free of child sexual abuse material where child safety is prioritized."

Derek Ray-Hill
Interim CEO,
Internet Watch Foundation

Conclusion

Collaborating to address emerging generative AI risks

We continue to invest in technologies that help protect and empower our users across all of Google's platforms, and we're also committed to working with stakeholders around the world to help advance smart and strong policy approaches for keeping young people safer online.

Together, we can build generative AI that is safer by design and contributes to a more secure online environment for everyone, especially children.