Microsoft

Addressing AI and child sexual exploitation and abuse risks: Microsoft's approach

Microsoft has a longstanding commitment to tackling child sexual exploitation and abuse (CSEA) risks on our services and to advancing a safer online ecosystem for children. From the landmark development of PhotoDNA technology in 2009, to our 2019 hackathon to tackle grooming, to the 2024 adoption of Thorn and All Tech is Human's **Safety by Design Principles** for AI, our approach continually evolves to respond to new risks and new technologies.

At Microsoft, we have taken a comprehensive approach to **addressing abusive AI-generated content**, including through a strong safety architecture grounded in safety by design, ongoing steps to safeguard our services from abusive content and conduct, and continued collaboration across industry and with governments and civil society. Our approach recognizes that generative AI's misuse poses serious child safety risks, and Microsoft has publicly pledged to adopt preventative and proactive measures across the AI lifecycle. To achieve this, we work with partners such as the National Center for Missing and Exploited Children (NCMEC), the Internet Watch Foundation, WeProtect Global Alliance, and the Tech Coalition to advance best practices and understand adversarial risks.

This report provides an overview of how we are embedding child safety safeguards into how we develop, deploy, and maintain AI systems, in keeping with the principles from Thorn and All Tech is Human. It should be read alongside Microsoft's **Responsible AI Transparency Report** and our **Digital Safety Content Report**.

Safety by design to prevent child sexual exploitation and abuse risks in our AI services

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

Microsoft ensures that its AI services are developed with stringent safety measures to address the risks of creating or facilitating child sexual abuse material. We implement robust safeguards to responsibly source and curate datasets and deploy safety systems.

- **Iterative red teaming and stress-testing:** Building AI safely is an ongoing process. Microsoft continuously improves our defenses by leveraging data from previous incidents to stress test content classifiers and strengthen our safeguards. This includes refining our risk categories and definitions, testing our mitigation layers using a robust evaluation set, and ensuring our safety systems adapt to real-world misuse scenarios while maintaining enterprise-grade trust and reliability. However, legal constraints prevent some red-teaming activities for CSEA, and we remain engaged with policymakers on this challenge.
- **Content provenance technology:** Microsoft is investing in content provenance solutions to aid in later identification of AI-generated content. For example, all AI images that are generated or modified using Azure OpenAI include Content Credentials, a secure, tamper-evident, and standardized way to add metadata that discloses the origin and history of content. Content Credentials are based on an open technical specification from the Coalition for **Content Provenance and Authenticity (C2PA)**, a Joint Development Foundation project.

Deploy: Safeguarding AI systems in deployment and use

Microsoft treats the deployment of generative AI as a critical juncture for safety and implements multiple protections in our AI products and services:

- **Layered content filtering and behavioral controls:** Microsoft generative AI products employ multi-layered filters and policies to prevent the generation of CSAM or other harmful content. Our default content filtering policies are designed to enforce protections against CSEA and establish a safety baseline across our deployments, including input filters and output classifiers to check prompts and responses. For example, if a user tries to prompt an image creator to produce sexual content involving a child, the input is blocked by the system before the image is generated. Likewise, if a prompt somehow bypassed input filtering, as can be the case in adversarial hacking scenarios, the model's output (image or text) is analyzed by dedicated classifiers for sexual and abusive content, which are intended to catch and suppress any CSEA depictions. We also enable **prompt transformation** for image generation, which acts as an additional safeguard against content harms. We also utilize metaprompts (also known as system messages)—guiding instructions to the model that instruct it to refuse requests for sexual content involving minors. Users are presented with a warning or a refusal message if they attempt disallowed queries, rather than any harmful content.

- **Clear policies and enforcement:** We are deeply committed to protecting children and ensuring that our AI systems are used responsibly, with safeguards in place to detect and block abusive behavior. Microsoft has clear terms of service and usage policies that forbid the misuse of our AI services to create illegal content. This includes an explicit prohibition on using our AI services (**enterprise** or **consumer**) to create or disseminate child sexual exploitation and abuse material or related activities that harm children. Microsoft has teams and automated systems monitoring for signals of attempted CSEA generation or other egregious abuse. We leverage PhotoDNA to mitigate the risk of CSEA being uploaded to Bing Image Creator and Microsoft Copilot. User attempts to circumvent our guardrails to generate or share illicit content may result in account suspension

and mandatory reporting to the National Center for Missing & Exploited Children. We have zero tolerance for any use of AI that facilitates child exploitation.

- **User feedback and reporting channels:** Alongside our preventive measures, Microsoft recognizes the importance of human feedback. We make it easy for users or third parties to report concerning outputs or misuse on our AI platforms. For instance, in Copilot, there is an in-product feedback button that enables users to report concerns. These reports are routed to trained moderators who review the prompt and output. Additionally, Microsoft operates a centralized "**Report A Concern**" webpage across our services, where the public can report suspected CSEA material.
  By incorporating user reporting loops, we add an extra safety net, and we will use any incidents to refine our safety measures.

- **Responsible hosting of third-party AI models:** We are equally committed to responsible hosting. For example, when we host third-party models, we make available Microsoft safety services and consider the extent to which they are configurable. If issues arise (say a third-party model is found producing unsafe outputs), Microsoft retains the right to remove or disable that model from our endpoints.

- **Educating and empowering developers:** Many of our customers will build their own applications on top of Microsoft AI models or APIs. We encourage shared ownership of safety by equipping these developers with guidance and tools. Alongside our AI services, we publish Transparency Notes and best practice documentation that highlight how to use the models responsibly and how to avoid potential misuse. We have also open-sourced various responsible AI tools—such as a version of our own red-teaming tool called PyRIT—to help others safety test their generative AI implementations.

Maintain: Ongoing monitoring, improvement, and collaboration

Advancing child online safety requires ongoing adaptation. Microsoft remains deeply committed to tackling CSEA harms across our services and to evolving our approach in response to new risks, regulations, and emerging best practices:

- **Continuous monitoring and incident response:** Microsoft monitors our AI services for abuse signals. We employ both automated systems and human moderators to identify patterns of potentially harmful behavior. For instance, telemetry may show if users are probing for disallowed content. We also stay alert to external reports—e.g., academic findings or other alerts about novel misuses of AI. If a new method to circumvent filters is discovered ("jailbreaking" prompts, etc.), we treat it as a trigger to improve our defenses, including rolling out updated content filter rules or model patches quickly.

- **Detecting and disrupting online CSEA:** As outlined in our Digital Safety Content Report, we have comprehensive **policies** prohibiting the use of our services to harm children. We prohibit CSAM and grooming children for sexual purposes, and our policies apply to content and conduct regardless of provenance. Whether synthetic or real, CSEA has no place on our services. We also prohibit the creation of such content, as well as sharing tools intended for its creation. We take a range of steps to enforce these policies across our services, including through the use of proactive detection technologies.

- **Investing in advanced research and tools:** The fight against child exploitation evolves as technology evolves. Microsoft collaborates with others in industry on innovation through the Tech Coalition, and we are a founding supporter of the Robust Open Online Safety Tools initiative (ROOST). Microsoft also works closely with OpenAI to share learnings, collaborate on mitigation strategies, and improve safety mechanisms. These lessons learned are considered when shipping model and system updates across the Azure OpenAI image stacks, ensuring improved defenses against CSEA-related content generation attempts by bad actors. These efforts help all users benefit from continuously enhanced protections against CSEA.

- **Transparency and progress reporting:** Microsoft has committed to regular public reporting, including through this paper, our Digital Safety Content Report, and our Responsible AI Transparency Report.