Amazon is committed to the following principles. We will:

DEVELOP: Develop, build and train generative AI models to proactively address child safety risks.

- **Responsibly source and safeguard our training datasets from CSAM and CSEM**: This is essential to helping prevent generative models from producing AIG-CSAM and CSEM. The presence of CSAM and CSEM in training datasets for generative models is one avenue in which these models are able to reproduce this type of abusive content. For some models, their compositional generalization capabilities further allow them to combine concepts (e.g. adult sexual content and non-sexual depictions of children) to then produce AIG-CSAM. We are committed to avoiding or mitigating training data with a known risk of containing CSAM and CSEM. We are committed to detecting and removing CSAM and CSEM from our training data, and reporting any confirmed CSAM to the relevant authorities. We are committed to addressing the risk of creating AIG-CSAM that is posed by having depictions of children alongside adult sexual content in our video, images and audio generation training datasets.
- **Incorporate feedback loops and iterative stress-testing strategies in our development process**: Continuous learning and testing to understand a model's capabilities to produce abusive content is key in effectively combating the adversarial misuse of these models downstream. If we don't stress test our models for these capabilities, bad actors will do so regardless. We are committed to conducting structured, scalable and consistent stress testing of our models throughout the development process for their capability to produce AIG-CSAM and CSEM within the bounds of law, and integrating these findings back into model training and development to improve safety assurance for our generative AI products and systems.
- **Employ content provenance with adversarial misuse in mind**: Bad actors use generative AI to create AIG-CSAM. This content is photorealistic, and can be produced at scale. Victim identification is already a needle in the haystack problem for law enforcement: sifting through huge amounts of content to find the child in active harm's way. The expanding prevalence of AIG-CSAM is growing that haystack even further. Content provenance solutions that can be used to reliably discern whether content is AI-generated will be crucial to effectively respond to AIG-CSAM. We are committed to developing state of the art media provenance or detection solutions for our tools that generate images and videos. We are committed to deploying solutions to address adversarial misuse, such as considering incorporating watermarking or other techniques that embed signals imperceptibly in the content as part of the image and video generation process, as technically feasible.

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

- **Safeguard our generative AI products and services from abusive content and conduct:** Our generative AI products and services empower our users to create and explore new horizons. These same users deserve to have that space of creation be free from fraud and abuse. We are committed to combating and responding to abusive content (CSAM, AIG-CSAM and CSEM) throughout our

generative AI systems, and incorporating prevention efforts. Our users' voices are key, and we are committed to incorporating user reporting or feedback options to empower these users to build freely on our platforms.

- **Responsibly host models:** As our models continue to achieve new capabilities and creative heights, a wide variety of deployment mechanisms manifests both opportunity and risk. Safety by design must encompass not just how our model is trained, but how our model is hosted. We are committed to responsible hosting of our generative models, assessing them e.g. via red teaming or phased deployment for their potential to generate AIG-CSAM and CSEM, and implementing mitigations before hosting. We are committed to responsibly hosting third party models in a way that minimizes the hosting of models that generate AIG-CSAM. We will ensure we have clear rules and policies around the prohibition of models that generate child safety violative content.
- **Encourage developer ownership in safety by design**: Developer creativity is the lifeblood of progress. This progress must come paired with a culture of ownership and responsibility. We encourage developer ownership in safety by design. We will endeavor to provide information about our models, including a child safety section detailing steps taken to avoid the downstream misuse of the model to further sexual harms against children. We are committed to supporting the developer ecosystem in their efforts to address child safety risks.

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

- **Prevent our services from scaling access to harmful tools:** Bad actors have built models specifically to produce AIG-CSAM, in some cases targeting specific children to produce AIG-CSAM depicting their likeness. They also have built services that are used to “nudify” content of children, creating new AIG-CSAM. This is a severe violation of children’s rights. We are committed to removing from our platforms and search results these models and services.
- **Invest in research and future technology solutions**: Combating child sexual abuse online is an ever-evolving threat, as bad actors adopt new technologies in their efforts. Effectively combating the misuse of generative AI to further child sexual abuse will require continued research to stay up to date with new harm vectors and threats. For example, new technology to protect user content from AI manipulation will be important to protecting children from online sexual abuse and exploitation. We are committed to investing in relevant research and technology development to address the use of generative AI for online child sexual abuse and exploitation. We will continuously seek to understand how our platforms, products and models are potentially being abused by bad actors. We are committed to maintaining the quality of our mitigations to meet and overcome the new avenues of misuse that may materialize.
- **Fight CSAM, AIG-CSAM and CSEM on our platforms**: We are committed to fighting CSAM online and preventing our platforms from being used to create, store, solicit or distribute this material. As new threat vectors emerge, we are committed to meeting this moment. We are committed to detecting and removing child safety violative content on our platforms. We are committed to disallowing and combating CSAM, AIG-CSAM and CSEM on our platforms, and combating fraudulent uses of generative AI to sexually harm children.

The detailed nature of specific mitigations that may be implemented to enact these principles are further recommended and defined in the associated whitepaper: Safety by Design for Generative AI: Preventing Child Sexual Abuse. More detail about the principles and the whitepaper can be found at https://teamthorn.co/gen-ai.

DEFINITIONS

AI-generated child sexual abuse material (AIG-CSAM): Visual depiction (image/video) of sexually explicit conduct involving a minor, the creation of which has been facilitated by generative AI technologies. This may range from a fully generated image/video to generated elements applied to a pre-existing image/video.

Child sexual abuse material (CSAM): Visual depiction (image/video) of sexually explicit conduct involving a minor. Does not require that the material depict a child engaging in sexual activity. Covers lewd and lascivious content, as well as content with a focus on genitalia. N.B. The definition of minor will vary depending on your legal jurisdiction.

Child sexual exploitation material (CSEM): Used as a shorthand for the full list of: image/video/audio content sexualizing children, grooming text, sexual extortion text, CSAM advertising, CSAM solicitation, and text promoting sexual interest in children.

CSAM advertising: Noting where child sexual abuse material can be found. This may be a URL or advertisement of CSAM for sale.

CSAM solicitation: The act of requesting, seeking out, or asking for access to, or the location of, child sexual abuse material.

Detect: The method or act of scanning through a larger set of data to attempt to identify the target material (e.g. CSAM or CSEM). Can include both manual and automated methodologies.