THORN

SAFETY BY DESIGN

Three-Month Progress Report

Report #3: November 2024 to January 2025

Author: Dr. Rebecca Portnoff

Companies' Commitment

All companies that agreed to commit to the Safety by Design principles also agreed to share the progress they have made in implementing those principles at a regular cadence. Thorn and All Tech Is Human recommended a quarterly cadence in all cases. Some companies agreed to quarterly reports, while others agreed to annual reports.

Every three months - Civitai, Invoke, and Metaphysic

Civitai and Metaphysic submitted their first report in July 2024, three months after their April 2024 commitments. Invoke joined the commitments in July 2024; its second report is included here, alongside the third round of reports from Civitai and Metaphysic.

Annually - Amazon, Anthropic, Google, Meta, Microsoft, Mistral AI, OpenAI, and Stability AI

Each of these companies joined the commitments in April 2024. Their first report will be included in the fourth round of reports from Civitai and Metaphysic, and the third report from Invoke.

As a result of the above, for this third public report, we focused our attention on Civitai (a platform for hosting third-party generative AI models), Invoke (a SaaS solution and OSS platform for AI image generation) and Metaphysic (a business that develops first-party generative AI models to create photorealistic generative AI video content for film studios).

Data Collection Process

To collect information about the progress it has made taking action on the Safety by Design principles, we sent each organization a survey. This survey requested information on both the steps it has taken in fulfillment of its commitments and the metrics used to measure the impact of its commitments. In

certain circumstances, we also conducted a follow-up interview to gather more detail on survey responses.

Below, we indicate how these companies have taken action on the principles based on their survey responses. We also provide analysis on what delta remains between the actions each company has taken and fulfilling the commitments it has made. It is worth noting that in line with the overarching "maintain" principles, though there currently may not be a delta between action taken and action needed, that is not a guarantee that there will not be in the future - as adversarial actors change their behaviors and a companies' tech stack develops and changes. Where we have the data, we include metrics to measure the impact of these actions to date.

This report documents the data self-reported by companies through the survey and any follow-up interviews. Thorn has not independently confirmed, investigated or audited the information provided in these self-reports. The data and this report are provided for general informational purposes. Thorn makes no representation or warranty of any kind, express or implied, regarding the accuracy, completeness or reliability of the data or the report, including the warranties of merchantability, fitness for a particular purpose, and non-infringement, and disclaims all liability related to creating, producing and issuing this report. All data provided to Thorn for this report is the property of the company providing such data and may be protected by applicable law. Links to third party websites are for informational purposes only, and the third party is responsible for the content on their website.

To read more about Thorn's strategy and perspective on accountability in regards to this Safety by Design initiative, see [1] in the references section at the end of this report.

Specific Findings

For a summary of their progress (as self-reported by the companies) and the action still needed to fulfill the commitments they have made, please see [1] in the references section at the end of this report. Below, we also provide a high-level overview (per sub-principle) for each company's current progress status. Progress is categorized as follows:

- **Not applicable:** According to the company, it does not currently offer a product or technology that fits the category of focus within the particular sub-principle. However, this can change as the company may expand its products and offerings.
- **Some progress:** Based on the company's self-reporting, it has made some progress in taking action on the particular sub-principle.
- **No current gaps observed:** Based on the company's self-reporting, it has met its commitments in taking action on the particular sub-principle. As noted above, while there may not currently be a delta between action taken and action needed, that is not a guarantee that there will not be in the future.
- **Not started:** Based on the company's self-reporting, it has not taken action on this particular sub-principle.

		Civitai	Invoke	Metaphysic
DEVELOP	Sub-principle 1	Not applicable	Not applicable	No current gaps observed
	Sub-principle 2	Not applicable	Not applicable	Not started
	Sub-principle 3	Not started	Some progress	Some progress
DEPLOY	Sub-principle 1	No current gaps observed	Some progress	No current gaps observed
	Sub-principle 2	Some progress	Not applicable	Some progress
	Sub-principle 3	Not started	Not applicable	No current gaps observed
MAINTAIN	Sub-principle 1	Some progress	No current gaps observed	Not applicable
	Sub-principle 2	No current gaps observed	No current gaps observed	No current gaps observed
	Sub-principle 3	Some progress	Some progress	Some progress

In sum, based on the company's self-reporting: Civitai has two sub-principles with no current gaps observed; three sub-principles where it has made some progress; two sub-principles it has not started; and two sub-principles that do not currently apply. Invoke has two sub-principles with no current gaps observed; three sub-principles where it has made some progress; zero sub-principles it has not started; and four sub-principles that do not currently apply. Metaphysic has four sub-principles with no current gaps observed; three sub-principles where it has made some progress; one sub-principle it has not started; and one sub-principle that does not currently apply.

Metaphysic indicated in its third report that it did not have any additional progress to report, since the second report. As a result, all sections for Metaphysic in this report remain as was previously documented in the second progress report.

PRINCIPLE 1

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

Sub-principle 1: Responsibly source and safeguard our training datasets from child sexual abuse material (CSAM) and child sexual exploitation material (CSEM).

This is essential to helping prevent generative models from producing AIG (AI generated)-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.

Civitai

CIVITAI REPORTS

According to Civitai, because it does not develop first-party generative AI models (it provides a platform for hosting of third-party generative AI models), it does not have any training datasets to curate or clean.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Civitai self-reports having implemented and what it committed to implementing.

Invoke

INVOKE REPORTS

According to Invoke, because it does not develop first-party generative AI models (it provides a SaaS solution and OSS platform for AI image generation), it does not have any training datasets to curate or clean.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports having implemented and what it committed to implementing.

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it has four primary strategies to enact this sub-principle. We address each of these strategies below:

1) Studio consent: According to Metaphysic, all data used for its generative AI models is sourced directly from the film studios with which it collaborates. As part of its contracts with these studios, Metaphysics reports that it requires the studio to warrant that no illegal material is present in these datasets.
2) User consent: According to Metaphysic, as part of its contracts with film studios it requires that studios also receive the consent of the individuals depicted in the data. It requires this consent for Metaphysic's use of both the data and its derivatives.
3) Human review: According to Metaphysic, upon receipt of the data, human moderators review every piece of data to confirm that no illegal or unethical content is present in the data.
4) Machine learning (ML)/AI dataset segmentation: According to Metaphysic, it uses proprietary ML/AI detection, to detect and separate out sexual content from depictions of children (such that its generative AI models are not trained on a combination of this content). Metaphysic reports its proprietary models for sexual content detection have an accuracy of around 95%. It reports more difficulty with the tools it uses for age estimation, with performance of these tools generally lower than the tools it uses for detecting sexual content.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Metaphysic self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Metaphysic reports the following metrics (as measured since joining the commitments):

- The percentage of its datasets that have been audited and updated: 100%.
- The number of instances of CSAM detected in its datasets: 0.
- The number of reports sent to NCMEC for CSAM and AIG-CSAM as a result of the above: 0 and 0.

Metaphysic further reports that in that process, it did not discover any CSAM in its datasets due to the nature of its business model as Metaphysic works exclusively with consensual data provided by clients and studios, and therefore has not submitted any reports to NCMEC or other reporting hotlines.

Sub-principle 2: 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.

Civitai

CIVITAI REPORTS

According to Civitai, because it does not develop first-party generative AI models (it provides a platform for hosting of third-party generative AI models), it does not have any first-party models to red team or otherwise stress test.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Civitai self-reports having implemented, and what it committed to implementing.

Invoke

INVOKE REPORTS

According to Invoke, because it does not develop first-party generative AI models (it provides a SaaS solution and OSS platform for AI image generation), it does not have any first-party models to red team or otherwise stress test. However, Invoke does report that it has performed red teaming exercises to test the robustness of its internal prompt monitoring solution, validating it in parallel with its previous prompt monitoring solution (askvera.io) before migration.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing.

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it has not yet incorporated consistent red teaming into its model development process due to its data governance model, which does not require emergency red teaming.

NOT YET IMPLEMENTED

Metaphysic will need to incorporate consistent red teaming for child safety violations in order to meet this commitment. The team chose to prioritize the work on data curation instead. The company has stated its intention to begin implementing consistent red teaming into its workflow in early 2025.

IMPACT METRICS

Metaphysic reports that it has conducted two red teaming exercises as "dry-runs" in advance of its planned implementation efforts in 2025.

Sub-principle 3: 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.

Civitai

CIVITAI REPORTS

In some cases, Civitai offers access to cloud-hosted third-party generative AI models on its platform. In these cases, Civitai has the necessary access to incorporate content provenance into the generated content (after generation). In cases where third-party generative AI models are cloud-hosted on Civitai's platform, Civitai reports it is actively exploring options to incorporate content provenance solutions post-generation, pending further industry standardization and technical developments.

NOT YET IMPLEMENTED

In order to meet this commitment, Civitai will need to ensure that the content generated by these cloud-hosted models include provenance information that is robust to adversarial misuse.

Invoke

INVOKE REPORTS

According to Invoke, all images created within its SaaS solution and OSS platform include metadata that contains a graph describing exactly how the image was created, along with other general

metadata about the image. Invoke further reports that this metadata is embedded within the image file itself, and can not be viewed by the majority of photo viewing applications, making it relatively difficult for the average user to remove or change the metadata.

Invoke reports that deciding what metadata to store within the images was a long process, and it continues to regularly assess and update that decision.

NOT YET IMPLEMENTED

Solutions that rely exclusively on metadata are vulnerable to adversarial misuse (e.g. metadata stripping). In order to meet this commitment, Invoke will need to assess the ways in which its current provenance solution is and is not robust to adversarial misuse, and, if necessary, support development and adoption of sufficiently robust solutions.

IMPACT METRICS

Invoke reports that 100% of images created within its SaaS solution and OSS platform include metadata describing the provenance of that image content.

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, C2PA is now implemented by default across its data pipelines.

NOT YET IMPLEMENTED

C2PA has built a strong technology foundation for companies to adopt. However, C2PA was not built with adversarial misuse in mind (e.g. it is vulnerable to metadata stripping). In order to meet this commitment, Metaphysic will need to engage with C2PA to better understand the ways in which C2PA is and is not robust to adversarial misuse, and, if necessary, support development and adoption of solutions that are sufficiently robust.

IMPACT METRICS

Metaphysic reports that 100% of its generative AI models have been developed with built-in content provenance.

PRINCIPLE 2

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

Sub-principle 1: 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.

Civitai

CIVITAI REPORTS

According to Civitai, it has four primary strategies to enact this sub-principle, for those cloud-hosted third-party generative AI models on its platform. We address each of these strategies below:

1) Detection at the inputs (i.e. where users submit prompts to the model): According to Civitai, these input-level detection defenses are a layered combination of automated filters and human review of content generation requests and subsequently generated media.

   Civitai reports that it combines keyword detection with ML/AI detection to identify prompts indicating an attempt to produce AIG-CSAM. Civitai's ML/AI prompt detection incorporates information from previous prompts submitted by users, to attempt to capture intent and broader context of the potentially violating prompt. Civitai further reports they are iterating on a new version of this system, and will have accuracy metrics to provide regarding the new system in time for the next report.

   According to Civitai, all prompts that are flagged by the automated filtering system are then sent to human review. For generated media that is confirmed by the human reviewer to be AIG-CSAM, a corresponding report is sent to NCMEC.

2) Enforcement at the outputs: Civitai reports using ML/AI detection to scan all cloud-model outputs for indications of minors, and sexually explicit or mature content. Civitai further reports that these efforts rely on in-house detection models, with reported accuracy rates of 75% to 80%. According to Civitai, all images that are flagged by the automated filtering system are then

sent to human review. For generated media that is confirmed by the human reviewer to be AIG-CSAM, a corresponding report is sent to NCMEC.

3) User reporting: According to Civitai, its users have the ability to report all uploaded content, including user accounts, models, model sample images, reviews, review images, comments, and outputs from cloud-hosted third-party models. Reported media items go into an internal queue for human review, where any verified CSAM and AIG-CSAM is then reported to NCMEC.

   According to Civitai, the reporting process for models and other users involves a longer form than the media report, requiring evidence of the violating behavior or capabilities (e.g. timestamps and metadata). For problematic models, a user report further requires evidence that the violative generated content was actually generated by the reported model itself. Civitai reports that once a model has been flagged as problematic, it is removed from access, and added to an internal Civitai hashlist such that future uploads of this same model are automatically blocked.

4) Prevention messaging: According to Civitai, when the automated filters detect that a user is attempting to prompt for AIG-CSAM, the user receives a real-time warning notification. Repeated attempts result in account suspension.

NOT YET IMPLEMENTED

1) Detection at the inputs: We currently do not see a gap between what Civitai self-reports having implemented, and what it committed to implementing.

2) Enforcement at the outputs: We currently do not see a gap between what Civitai self-reports having implemented, and what it committed to implementing.

3) User reporting: We currently do not see a gap between what Civitai self-reports having implemented, and what it committed to implementing.

4) Prevention messaging: We currently do not see a gap between what Civitai self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Civitai reports the following metrics (as measured since joining into the commitments):

- The number of violative prompts detected at the inputs: 252,511[1]
- The number of user reports submitted for various model violations: 11,048
- The number of prevention messages surfaced due to violative prompts: 1,262,555[2]
- The number of instances of AIG-CSAM detected at the outputs: 178
- The number of reports sent to NCMEC for CSAM and AIG-CSAM as a result of the above: N/A and 178

[1] Civitai reports that this represents 0.027% of all requests
[2] Civitai reports that this number is higher than the number of violative prompts detected, because Civitai surfaces prevention messaging earlier in its overall process of establishing intent and broader context of potentially violating prompts

Invoke

INVOKE REPORTS

According to Invoke, it has four primary strategies to enact this sub-principle, for its SaaS solution. We address each of these strategies below:

1) Detection at the inputs (i.e. where users submit prompts to the model): According to Invoke, its input-level detection defenses are implemented via prompt monitoring, such that Invoke can detect, ban, and report any users attempting to create abusive content on its hosted products. Invoke further reports that it has migrated its input-level prompt monitoring detection to a self-managed solution for detecting abusive inputs. According to Invoke, whenever violations of acceptable use are detected on its platform, it regularly warns, bans, and reports users based on the severity of the attempted generation. Invoke reports that its detection solution errs on the side of false positives vs. false negatives, as the company has not yet identified a case where the solution has missed abusive inputs such that the user inputting the problematic inputs was not reported. Invoke further reports that it has implemented more rigorous fingerprinting and blocking to prevent abusive users who have already been banned from accessing the platform through secondary or alternative accounts.

   According to Invoke, it commits time every day to monitoring the actions detected by the above measures to review and respond to them accordingly.

2) Customer feedback: According to Invoke, it has existing workflows to allow for customer feedback on any and all issues related to the generated media its SaaS solution customers produce using its platform, including any feedback related to content that may contain illegal or unethical material. Invoke further reports that it has published resources for reporting abusive content found, and invited users with concerns to reach out to Invoke's support team with additional details where necessary.

3) Prevention messaging: Invoke reports that when a user is detected attempting to use a model that has been optimized for the creation of AIG-CSAM (e.g. fine-tuned on CSAM) for CSAM, Invoke will subsequently direct the user to redirectionprogram.com.

4) Model suppression: Invoke reports that it makes use of Thorn's hashlist of models, where the models on that list have been verified as having been optimized for the creation of AIG-CSAM (e.g. fine-tuned on CSAM). According to Invoke, it uses this hashlist to ensure that all uploads of these models are automatically blocked.

In regards to its OSS offering, Invoke reports that it has two primary strategies to enact this sub-principle: prevention messaging and model suppression. According to Invoke, both of these strategies are implemented in its OSS offering, in the same way as they are implemented for its SaaS solution. Invoke further notes that for its OSS offering, it has found that this form of open source deployment both allows its business to receive far more QA and testing than may be the norm within its field, but also results in its services being leveraged in ways it cannot fully control.

NOT YET IMPLEMENTED

For Invoke's SaaS solution:

1) Detection at the inputs: Invoke is not yet detecting at the inputs for CSAM (e.g. via using hashing and matching against verified CSAM lists to detect known CSAM as part of input-level detection defenses of its SaaS solution.) In order to meet this commitment, Invoke will need to incorporate detection at the inputs for CSAM images provided as input to generative AI models and systems.
2) Enforcement at the outputs: Invoke does not currently have a system in place for automated filtering and comprehensive human review as part of its output-level detection defenses of its SaaS solution. In order to meet this commitment, Invoke will need to incorporate detection at the outputs into its SaaS solution moderation system.
3) Customer feedback: We currently do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing.
4) Prevention messaging: We currently do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing.
5) Model suppression: We currently do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing.

For Invoke's OSS offering, it has not yet incorporated user reporting functionality to allow users of its OSS offering to report violative models it has discovered (e.g. models that can produce AIG-CSAM, or were otherwise optimized for producing AIG-CSAM), to the appropriate organizations. Invoke also has not incorporated prevention messaging for input prompts attempting to generate AIG-CSAM via its OSS offering. Invoke reports it will continue to evaluate ways it can improve its prevention strategy long-term. In order to meet this commitment, Invoke will need to iterate and expand on its user reporting and prevention strategy for its OSS offering.

IMPACT METRICS

Invoke reports the following metrics (as measured since joining into the commitments):

- The number of violative prompts detected at the inputs: 2822.
- The number of user reports submitted for various model violations: 0.
- The number of reports sent to NCMEC for CSAM and AIG-CSAM as a result of the above: N/A and 2822.

According to Invoke, all of the above metrics are sourced from its SaaS solution, as Invoke does not have telemetry or access to collect metrics for its OSS platform. Invoke further noted that it did not expect the frequency at which people would attempt to perform such actions on a commercially hosted product.

According to Invoke, it does not have any metrics to report for the following items, as it has not yet implemented the underlying interventions:

- The number of instances of CSAM detected at the inputs
- The number of instances of AIG-CSAM detected at the outputs

- The number of prevention messages surfaced due to violative prompts

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it has three primary strategies to enact this sub-principle. We address each of these strategies below:

1) Controlled access: According to Metaphysic, no one outside of the employees at Metaphysic has access to its generative AI models. Film studios only receive the requested outputs that they have contracted with Metaphysic to produce. This is part of Metaphysic's larger strategy to ensure that, from a business and ethics perspective, the generative AI models it builds are only used to generate content for the specific use case in which it has been contracted.
2) Human moderation: As noted in the analysis on the sub-principle "Responsibly source and safeguard our training datasets from CSAM and CSEM," Metaphysic reports that it employs human moderators to review every piece of received film studio data for illegal and unethical content. Metaphysic similarly reports employing human moderators to review every piece of generated media for the same purpose.
3) Customer feedback: According to Metaphysic, it has existing workflows to allow for customer feedback on any and all issues related to the generated media it produces for its customers, including any feedback related to content that may contain illegal or unethical material.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Metaphysic self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Metaphysic reports the following metrics (as measured since joining into the commitments):

- The number of instances of CSAM detected at the inputs: 0.
- The number of user reports submitted for various violations: 0.
- The number of instances of AIG-CSAM detected at the outputs: 0.
- The number of reports sent to NCMEC for CSAM and AIG-CSAM as a result of the above: 0 and 0.

Metaphysic reports that with the above strategies in place, it has not discovered any CSAM or AIG-CSAM produced by its generative AI models, and therefore has not submitted any reports to NCMEC or other reporting hotlines.

Sub-principle 2: Responsibly host our 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 first party 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 also 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.

Civitai

CIVITAI REPORTS

Civitai reports that it has established terms of service that prohibit the use and upload of third-party generative AI models on its platform for generating AIG-CSAM, sexually exploitative depictions of children, or photorealistic depictions of minors. According to Civitai, it enforces these policies by employing a combination of human moderation and automated review. Civitai reports they use a combination of in-house and external solutions (specifically, Hive's Visual Moderation API and Hive's Demographic API) for the automated review, such that the titles, descriptions, and images associated with the generative model are assessed for presence of minors. In a final pass, these predicted labels are combined with (where relevant) the prompt via their automated review system.

Civitai further reports that when violative models are identified through user reporting, Civitai takes action by either removing them from the platform (see "User reporting" in the "Safeguard our generative AI products and services from abusive content and conduct" sub-principle above) or implementing mitigations to prevent misuse. According to Civitai, it utilizes semi-permeable membranes [2] (SPMs) to ensure that cloud-hosted generative AI models do not produce AIG-CSAM or other harmful content as part of its proactive safety measures. Civitai further reports that certain models are restricted to cloud-hosted generation only, where filtering and SPM mitigations can be applied effectively.

NOT YET IMPLEMENTED

Civitai has not yet incorporated mitigations for all of the Stable Diffusion 1.5 models (and its derivatives) hosted on its site. This model and its derivatives have been confirmed as capable of generating AIG-CSAM [3]. According to Civitai, it is actively working toward incorporating mitigations for Stable Diffusion 1.5 models and its derivatives hosted on its platform.

Civitai reports that it has successfully implemented SPM mitigation in its cloud-hosted generative AI models and it now aims to extend these safeguards to all hosted models on its platform, to ensure a consistent and effective approach to mitigating potential risks. Civitai reports it has evaluated the integration of its SPM technology in these additional models to enhance safeguards and eliminate the capability to produce AIG-CSAM. According to Civitai, its current approach has been implemented

on a single series of models and testing is being done to expand the safety enhancement to the entirety of the Stable Diffusion 1.5 ecosystem of models. To meet this commitment, Civitai will need to ensure that the models available for download from its site incorporate these same mitigations where necessary.

Civitai is not yet assessing newly uploaded generative AI models (before they are uploaded to the platform) for their capability to generate child safety violative content. It is also not comprehensively retroactively assessing its currently hosted models. Civitai reports that identifying and mitigating third-party models capable of generating child safety violative content remains a complex and evolving challenge. According to Civitai, it is actively exploring methods to assess newly uploaded generative AI models for potential child safety risks before they become publicly accessible. Civitai reports that while comprehensive retroactive assessments of hosted models remain a challenge due to the lack of automated model assessment technology, it has conducted early research evaluating possible scalable approaches - such as leveraging a dedicated GPU cluster to test models with predefined prompts and automated ML/AI detection at the outputs. Civitai reports that while current hardware limitations prevent full-scale implementation, it anticipates launching a beta system later this year.

Civitai further reports it has explored additional methods to enhance its assessment of generative AI models before they are hosted on the platform. One potential approach it reports exploring involves leveraging existing metadata and categorization tags, as existing research [5] indicates a correlation between "Deepfakes" and "NSFW content" tags on Civitai's platform, highlighting the potential for automated tagging to aid in safety assessments. Civitai reports that metadata alone is insufficient for a proper evaluation, but may be useful as an additional source of information for evaluation.

Similarly, information collected in child safety sections of a model card (detailing steps taken to mitigate for harms) could also open the ability to incorporate a layer of assessment focused on the child safety section of the model card, before allowing those models to be hosted. For example, Civitai could include questions to third-party developers on what technologies were used to implement data cleaning and curation. Answers with insufficient detail, or other indications that the provided response is false, could result in Civitai disallowing the model to be hosted on its platform.

To meet its commitments, Civitai will need to incorporate systematic model assessment of the third-party generative AI models hosted on its platform, for their capability to produce AIG-CSAM and other child safety violative content.

IMPACT METRICS

Civitai reports the following metrics (as measured since joining into the commitments):

- The number of hosted generative AI models taken down and removed from platform access, due to discovering they are capable of producing AIG-CSAM and CSEM: 198
- Of those models, the number of models for which mitigations were incorporated and the model was re-uploaded: 15

According to Civitai, it does not have any metrics to report for the following item, as it has not yet implemented the underlying interventions:

- The percentage of newly hosted generative AI models that have been assessed for their ability to produce AIG-CSAM and CSEM before being made accessible

Invoke

INVOKE REPORTS

According to Invoke, it does not serve as a platform for third-party developers to distribute or merchandise their models, nor does it build any first-party models. However, Invoke reports it has proactively established terms of service that prohibit customer use of Invoke's services in a way that violates any law, regulation or court order, including the use of third-party generative AI models (within its SaaS solution and OSS systems) to generate AIG-CSAM and other sexually exploitative depictions of children. Invoke further reports it has established user policies and enforcement mechanisms around the upload and subsequent use of models that are capable of generating AIG-CSAM (such as Stable Diffusion 1.5 models and its derivatives), as noted in the discussion around the principle "Safeguard our generative AI products and services from abusive content and conduct."

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing.

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it does not host any third-party models. However, when making use of third-party models internally, Metaphysic reports that it assesses every model, prior to using said model, for a variety of ethical issues (including child safety violations). If any issues are found, Metaphysic does not use the model.

When considering the first-party models it builds, as noted in the discussion for the sub-principle "Incorporate feedback loops and iterative stress-testing strategies in our development process," Metaphysic reports it has not yet incorporated red teaming into its processes. However, Metaphysic does report that it practices model assessment and phased deployment of its models. According to Metaphysic, this model assessment is currently manual. Metaphysic reports it is working towards solutions to conduct these assessments systematically and in an automated fashion, but doing so requires significant resources to build. Finally, as noted in the discussion on "Safeguard our generative AI products and services from abusive content and conduct," Metaphysic reports that no individuals or organizations outside of Metaphysic have direct access to its generative AI models.

NOT YET IMPLEMENTED

To meet its commitments, Metaphysic will need to incorporate systematic model assessment of its generative AI models for their capability to produce AIG-CSAM and other child safety violative content.

IMPACT METRICS

Metaphysic reports that 100% of its first-party models undergo phased deployment.

Sub-principle 3: 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.

Civitai

CIVITAI REPORTS

Civitai reports that it has not made progress to include a child safety section in the model card equivalent (i.e. the model "details" section) for third-party model developers to fill in before they upload their model.

NOT YET IMPLEMENTED

Civitai reports it is actively evaluating ways to enhance transparency and safety in model submissions. According to Civitai, one approach under consideration is requiring developers to confirm that their training data has been properly curated and cleaned, in line with the "Develop" sub-principle "Responsibly source and safeguard our training datasets from CSAM and CSEM." Civitai further reports it is mindful of balancing the level of detail included, ensuring that the information enhances safety without inadvertently guiding bad actors toward models lacking proper safeguards.

As part of this effort, Civitai reports it is also exploring how submitted safety information could be used as a factor in approving models for hosting, aligning with its broader commitments to responsible deployment as stated in the "Deploy" sub-principle "Responsibly host our models."

To meet this commitment, Civitai will need to update its model card equivalents to include a child safety section detailing steps the third-party model developer has taken to follow the "Develop" principles, or implement other equivalent strategies to encourage the third-party model developer to address child safety risks.

Invoke

INVOKE REPORTS

Invoke reports that it does not develop first-party models, nor does it serve as a platform for third-party developers to distribute or merchandise their models, and therefore it does not make use of model cards in either capacity. In regards to the third-party models that are uploaded to its SaaS solution or OSS offerings, according to Invoke it offers a "Name" and "Description" field to customers, where customers can choose to input details regarding the third-party model they are using.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing. However, there is an opportunity here for Invoke to point customers towards other existing documentation (e.g. model cards included on model hosting platforms) that provide more context and relevant information to its customers regarding what child safety interventions were put into place as part of model development.

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it has incorporated into its datasets and models an associated card. Metaphysic reports that this card contains information listed in the “Model Card: Child Safety” additional resource included in [6].

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Metaphysic self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Metaphysic reports that 100% of its datasets and models have the above described card implemented.

PRINCIPLE 3

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

Sub-principle 1: 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.

Civitai

CIVITAI REPORTS

According to Civitai, known and verified problematic models (discovered via user reporting) are removed from access, and added to an internal Civitai hashlist such that future uploads of this same model are automatically blocked. Additionally, Civitai reports it retroactively checks its existing corpus of hosted models, running a daily batch job to detect and remove said models such that the newly discovered problematic models do not appear anywhere else in its collection.

Civitai further reports it has updated its policies such that any AI workflows, models, or tools designed with the intention of removing clothing or otherwise "nudifying" individuals in any context ("real people" or otherwise) is explicitly prohibited. According to Civitai, it has incorporated periodic manual moderation efforts to enforce these policies, where moderators will search for indicators in the title and description of resources and workflow that indicate they violate Civitai's policies regarding nudifying individuals.

Additionally, Civitai reports that its existing policies against "suggestive" or "sexual" content depicting real people, combined with its use of prompt filters and SPMs for cloud-generated images captures a significant portion of nudifying activity.

NOT YET IMPLEMENTED

Civitai has not yet incorporated automated efforts for enforcement of its policies regarding nudifying AI workflows, models and tools. To meet its commitment on this sub-principle, Civitai will need to update its enforcement mechanisms to incorporate automated strategies for detection of these services, such that these services (that, regardless of how they are advertised, e.g. for use on children, for use on adults, for use on real people vs. fictional characters, can and are being used to nudify children [7]) are removed.

IMPACT METRICS

Civitai reports the following metrics (as measured since joining into the commitments):

- The number of models optimized to produce AIG-CSAM, retroactively removed: 183
- The number of prevented attempts to upload a model optimized to produce AIG-CSAM: Civitai reports that it is updating its tracking mechanism, and will be able to provide this information again for the next report.
- The number of nudifying services, retroactively removed: 4

According to Civitai, it does not have any metrics to report for the following item, as it has not yet implemented the underlying interventions:

- The number of prevented attempts to upload a nudifying model or nudifying workflow

Invoke

INVOKE REPORTS

According to Invoke, it makes use of Thorn's hashlist of models, where the models on that list have been verified as having been optimized for the creation of AIG-CSAM (e.g. fine-tuned on CSAM). According to Invoke, it uses this hashlist to ensure that all uploads of these models (in both its SaaS solution and its OSS offering) are automatically blocked (see "Model suppression" in the "Safeguard our generative AI products and services from abusive content and conduct" sub-principle above). Invoke further reports that it retroactively checks uploaded models in its SaaS solution, when new models are added to Thorn's hashlist of models.

In regards to the upload and use of models that are intended for "nudifying" imagery, Invoke reports that it does not have the necessary contextual information (e.g. the advertising language used by the provider of the model indicating it is a "nudifying" model) to reliably distinguish between a model that has been built for the express purpose of "nudifying" imagery, vs. models that are capable of nudifying imagery but were not built for that express purpose. As a result, Invoke reports that the user policies and enforcement mechanisms noted in the "Safeguard our generative AI products and services from abusive content and conduct" sub-principle outline its strategy for addressing this type of misuse.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Invoke reports the following metrics (as measured since joining into the commitments):

- The number of prevented attempts to upload a model optimized to produce AIG-CSAM: 0.
- The number of models optimized to produce AIG-CSAM, retroactively removed: 0.

According to Invoke, the above metrics are sourced from its SaaS solution offering, as Invoke does not have telemetry or access to collect metrics for its OSS platform. Invoke further notes that it has never had a user attempt to upload to its SaaS solution system a model from Thorn's hashlist of

models, where the models on that list have been verified as having been optimized for the creation of AIG-CSAM (e.g. fine-tuned on CSAM).

According to Invoke, it does not have any metrics to report for the following items, as it does not have the necessary contextual information to reliably distinguish between a model that has been built for the express purpose of "nudifying" imagery, vs. models that are capable of nudifying imagery but were not built for that express purpose:

- The number of nudifying models, retroactively removed
- The number of prevented attempts to upload a nudifying model

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it does not host third-party models or services, or offer search functionality as part of its business model.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Metaphysic self-reports having implemented, and what it committed to implementing.

Sub-principle 2: 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.

Civitai

CIVITAI REPORTS

According to Civitai, it has invested in and deployed future technology solutions via its line of work around SPM-based interventions. Civitai further reports that it monitors its user community for emerging risks, and relies on outside partners to also monitor trends and emerging risks. Additionally, Civitai reports continuous effort improving the ML/AI detection technology it builds in-house.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Civitai self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Civitai reports the following metrics (as measured since joining into the commitments):

- How much investment (e.g. R&D time, grants/funding, etc.) has been made into tools to protect content from AI-generated manipulation: ~25% of all development time has been spent optimizing and improving moderation tools, accounting for nearly 20% of payroll costs during that time frame. In the last quarter, Civitai hired 2 Contractors and 1 FTE to support these efforts.
- Cadence at which mitigations are assessed against the business' tech stack, to ensure effective performance: Quarterly.

Invoke

INVOKE REPORTS

According to Invoke, it has invested in and deployed future technology solutions via its work building its own detection mechanisms and systems. Invoke further reports it has developed new checks for blocked accounts based on card fingerprints from payment processors to prevent repeat abusers from accessing its platform.

In addition, Invoke reports that it leverages its access to OSINT using forums such as Github, Reddit, Discord etc. to monitor for emerging risks. Invoke further reports that all new features created on its platform are architected with the explicit goal of avoiding the creation of abusive content in mind.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Invoke reports the following metrics (as measured since joining into the commitments):

- How much investment (e.g. R&D time, grants/funding, etc.) has been made into tools to protect content from AI-generated manipulation: $224,000 in R&D time and tools.
- Cadence at which mitigations are assessed against the business' tech stack, to ensure effective performance: Multiple times a week.

Metaphysic

METAPHYSIC REPORTS

As noted in the discussion on "Safeguard our generative AI products and services from abusive content and conduct," according to Metaphysic no individuals or organizations outside of Metaphysic have direct access to its generative AI models. As a result of this controlled access, Metaphysic reports it has not made use of open source intelligence (OSINT) or other strategies to understand

how bad actors are potentially misusing its products and services. In regards to investing in research and technology, Metaphysic reports that it intends to publish its findings around its efforts to build ML/AI dataset segmentation technologies. Metaphysic further reports (as outlined in the discussion on "Responsibly host our models") its investment in building scalable, automated model assessment mechanisms.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Metaphysic self-reports having implemented, and what it committed to implementing.

IMPACT METRICS

Metaphysic reports the following metrics (as measured since joining into the commitments):

- How much investment (e.g. R&D time, grants/funding, etc.) has been made into tools to protect content from AI-generated manipulation: Metaphysic cannot disclose this figure.
- Cadence at which mitigations are assessed against the business' tech stack, to ensure effective performance: Once per month.

Sub-principle 3: 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.

Civitai

CIVITAI REPORTS

According to Civitai, it employs a multi-layered approach to safeguarding its platform, utilizing the same core strategies outlined in the "Safeguard our generative AI products and services from abusive content and conduct" sub-principle: detection, user reporting, and prevention messaging.

In addition to using the in-house detection models discussed in the previously mentioned sub-principle, Civitai reports that when conducting ML/AI detection to scan uploaded images for indications of minors, sexually explicit or mature content, it also leverages external tools such as Hive moderation. Civitai further reports it maintains an internal hash database of removed images to prevent the re-upload of previously flagged content, ensuring that identified violations do not resurface. Additionally, Civitai reports it detects uploads of images depicting known, real humans (in order to prevent sexual deepfakes of known individuals) checking input images against an unspecified database of "known individuals".

Civitai further reports that it ensures that reports of AIG-CSAM submitted to NCMEC's CyberTipline include all necessary parameters for accurate reporting and intervention, inclusive of information regarding the model used to generate the offending image, when that information is known.

NOT YET IMPLEMENTED

Civitai is not yet using hashing and matching against third-party owned, maintained and verified CSAM lists to detect known CSAM hosted on its platform. According to Civitai, it is working to expand its moderation capabilities by integrating additional industry-standard tools. Civitai reports it is actively pursuing access to Microsoft's pDNA license, which would allow for integration with NCMEC's verified CSAM hashlist.

Civitai does not yet employ prevention messaging as part of safeguarding the search functionality on its site (e.g. entering the terms "child abuse model" into its in-site search functionality does not surface prevention messaging). Civitai reports it is exploring improvements to its search functionality to incorporate prevention messaging to ensure that certain flagged search terms trigger warnings or deterrent messaging.

To meet its commitment, Civitai will need to incorporate hashing and matching against verified CSAM lists in its overall content moderation strategy, as well as incorporate prevention messaging for the search functionality on its site.

IMPACT METRICS

Civitai reports the following metrics (as measured since joining into the commitments):

- The number of instances of CSAM detected on its site: 61
- The number of user reports submitted for various violations on its site: 560,555
- The number of instances of AIG-CSAM detected on its site: 178
- The number of reports sent to NCMEC for CSAM and AIG-CSAM as a result of the above: 61 and 178

Civitai further reports that as a result of these various violations, 17,436 accounts have been banned.

According to Civitai, it does not have any metrics to report for the following items, as it has not yet implemented the underlying interventions:

- The number of prevention messages surfaced

Invoke

INVOKE REPORTS

According to Invoke, the strategies outlined in "Safeguard our generative AI products and services from abusive content and conduct," are comprehensive across its SaaS solution and OSS offerings. According to Invoke, when reporting AIG-CSAM to NCMEC its content moderation team ensures that its CyberTipline reports supply all of the correct parameters.

NOT YET IMPLEMENTED

As noted in the discussion for the sub-principle "Safeguard our generative AI products and services from abusive content and conduct," Invoke is not yet detecting CSAM uploaded to its SaaS solution system. This intervention is applicable both for detection at the inputs (as discussed previously) and for detection with user datasets that are uploaded to Invoke's SaaS solution offering for training and fine-tuning customer models. To meet its commitment, Invoke will need to incorporate CSAM detection in its overall content moderation strategy.

IMPACT METRICS

According to Invoke, it does not have any metrics to report for the following items, as it has not yet implemented the underlying interventions:

- The number of instances of CSAM detected in user datasets
- The number of reports sent to NCMEC for CSAM as a result of the above

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, it does not build or offer access to platforms that allow for the solicitation or distribution of any material (regardless of the type of material that is solicited or distributed). In regards to preventing the creation and storing of this material, see the discussion around the principle "Develop, build and train generative AI models that proactively address child safety risks."

NOT YET IMPLEMENTED

For more detail on progress, please see the discussion around the principle "Develop, build and train generative AI models that proactively address child safety risks."

IMPACT METRICS

For more detail on impact metrics, please see the discussion in previous principles.

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.

References

1. Thorn. "Thorn's Safety by Design for Generative AI: Progress Reports." *Thorn*, 20 March 2025, https://www.thorn.org/blog/thorns-safety-by-design-for-generative-ai-progress-reports.
2. Lyu, Mengyao, et al. One-Dimensional Adapter to Rule Them All: Concepts, Diffusion Models and Erasing Applications. *CVPR, 2024.* https://doi.org/10.48550/arXiv.2312.16145.
3. Thiel, D., Stroebel, M., and Portnoff, R. (2023) Generative ML and CSAM: Implications and Mitigations. Stanford Digital Repository. Available at https://doi.org/10.25740/jv206yg3793.

4. Thiel, D. (2023). Identifying and Eliminating CSAM in Generative ML Training Data and Models. Stanford Digital Repository. Available at https://doi.org/10.25740/kh752sm9123.

5. Wei, Yiluo, et al. Exploring the Use of Abusive Generative AI Models on Civitai. *ACM Multimedia 2024.* https://doi.org/10.48550/arXiv.2407.12876.

6. Thorn and ATIH. (2024) Safety by Design for Generative AI: Preventing Child Sexual Abuse. Thorn Repository. Available at https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf.

7. *Western District of North Carolina | Charlotte Child Psychiatrist Is Sentenced To 40 Years In Prison For Sexual Exploitation of A Minor And Using Artificial Intelligence To Create Child Pornography Images Of Minors | United States Department of Justice*. 8 Nov. 2023, https://www.justice.gov/usao-wdnc/pr/charlotte-child-psychiatrist-sentenced-40-years-prison-sexual-exploitation-minor-and.