THORN

SAFETY BY DESIGN

Three-Month Progress Report

Report #2: August 2024 to October 2024

Author: Dr. Rebecca Portnoff

Companies' Commitment

All companies that agreed to commit to the Safety by Design principles also agreed to commit to sharing progress they have made taking action on those principles, at a regular cadence. In all cases, Thorn and All Tech Is Human recommended a quarterly cadence. 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 of 2024, three months after their April 2024 commitments. Invoke joined the commitments in July 2024; their first report is included here, alongside the second 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 come in with the fourth round of reports from Civitai and Metaphysic, and the third report from Invoke.

As a result of the above, for this second 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 they have made taking action on the Safety by Design principles, we sent each organization a survey. This survey requested information both on what steps they have taken in fulfillment of their commitments, as well as metrics to measure the impact of their

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 action they have taken and fulfilling the commitments they have 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 companies' current progress status. Progress is categorized as followed:

- **Not applicable:** According to the company, they do not currently offer a product or technology that fits the category of focus within the particular sub-principle. This can change, as the company may grow its products and offerings.
- **Some progress:** Based on the company's self-reporting, they have made some progress in taking action on the particular sub-principle.
- **No current gaps observed:** Based on the company's self-reporting, they have met their commitments in taking action on the particular sub-principle. As noted above, while 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.
- **Not started:** Based on the company's self-reporting, they have not made progress in taking 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	Some progress	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 one sub-principle with no current gaps observed; four sub-principles where they have made some progress; two sub-principle they have 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 they have made some progress; zero sub-principles they have 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 they have made some progress; one sub-principle they have not started; and one sub-principle that does not currently apply.

Both Civitai and Metaphysic indicated in their second report that they did not have any additional progress to report, since the first report. When asked to provide context into this, Metaphysic reported that they are in the middle of a funding process, and therefore resources were allocated to other important tasks in the business. Civitai similarly indicated other work had taken priority. As a result, the only sections in this progress report that have been updated in regards to Civitai and Metaphysic are in their reported metrics. All other sections for Civitai and Metaphysic in this report remain as was previously documented in the first 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 they do not develop first-party generative AI models (they provide a platform for hosting of third-party generative AI models), they do not have any training datasets to curate or clean.

NOT YET IMPLEMENTED

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

Invoke

INVOKE REPORTS

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

NOT YET IMPLEMENTED

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

Metaphysic

METAPHYSIC REPORTS

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

1) Studio consent: According to Metaphysic, all data they use for their generative AI models is sourced directly from the film studios with which they collaborate. As part of their contracts with these studios, Metaphysics reports they require that the studio warrant that no illegal material is present in these datasets.
2) User consent: According to Metaphysic, as part of their contracts with film studios they require that studios also receive the consent of the individuals depicted in the data. They require 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) ML/AI dataset segmentation: According to Metaphysic, they use proprietary ML/AI detection, to detect and separate out sexual content from depictions of children (such that their generative AI models are not trained on a combination of this content). Metaphysic reports their proprietary models for sexual content detection have an accuracy of around 95%. They report more difficulty with the tools they use for age estimation, with performance of these tools generally lower than the tools they use for detecting sexual content.

NOT YET IMPLEMENTED

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

IMPACT METRICS

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

- The percentage of their datasets that have been audited and updated: 100%.
- The number of instances of CSAM detected in their datasets: 0.
- The number of reports sent to NCMEC for CSAM and AIG-CSAM: 0 and 0.

Metaphysic further report that in that process, they did not discover any CSAM in their datasets due to the nature of their business model as Metaphysic works exclusively with consensual data provided by clients and studios, and therefore have 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 they do not develop first-party generative AI models (they provide a platform for hosting of third-party generative AI models), they do 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 as having implemented, and what they committed to implementing.

Invoke

INVOKE REPORTS

According to Invoke, because they do not develop first-party generative AI models (they provide a SaaS solution and OSS platform for AI image generation), they do not have any first-party models to red team or otherwise stress test. However, Invoke does report that they have performed red teaming exercises to test the robustness of their prompt monitoring solution (askvera.io).

NOT YET IMPLEMENTED

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

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, they have not yet incorporated consistent red teaming into their model development process, due to their data governance model that 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. They have stated their intention to begin implementing consistent red teaming into their workflow in early 2025.

IMPACT METRICS

Metaphysic reports that they have conducted two red teaming exercises as "dry-runs" in advance of their 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 third-party models on their platform which are cloud-hosted. In these cases, Civitai has the necessary access to incorporate content provenance into the generated content (after generation). In cases where third-party models are cloud-hosted on Civitai's platform, Civitai reports they are 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 their 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 they continue 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 their current provenance solution is and is not robust to adversarial misuse, and - if necessary - support development and adoption of solutions that are sufficiently robust.

IMPACT METRICS

Invoke reports that 100% of images created within their 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 their 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 their 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, they have three primary strategies to enact this sub-principle, for those third-party models on their platform which are cloud-hosted. 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 media inputs.

   Civitai reports that they combine keyword detection with ML/AI detection to identify prompts indicating an attempt to produce AIG-CSAM. Their 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. Additionally, Civitai reports they maintain an internal hash database of previously removed images to prevent re-upload.

   Civitai further reports using ML/AI detection to scan any input images for indication of minors, and sexually explicit or mature content. According to Civitai, they also detect 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”. For image detection, they report using a combination of external ML/AI detection models (Amazon Rekognition) and models built in-house. Civitai reports their in-house models have an accuracy of around 75% to 80%.

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

2) User reporting: According to Civitai, 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.

3) 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: Civitai is not yet using hashing and matching against verified CSAM lists to detect known CSAM at the inputs of their systems. They report that they are currently attempting to get access to Microsoft's pDNA license such that they would be able to begin hashing/matching against NCMEC's verified CSAM hashlist. In order to meet this commitment, Civitai will need to incorporate additional hashing and matching against industry-shared verified CSAM lists into their layered moderation system.

2) Enforcement at the outputs: Civitai does not currently have a system in place for automated filtering and comprehensive human review at the outputs of their cloud-hosted third-party models. In order to meet this commitment, Civitai will need to incorporate detection at the outputs of their cloud-hosted generative AI models.

3) Prevention messaging: We currently do not see a gap between what Civitai self-reports as having implemented, and what they 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: >120,000 prompts. 100,000 were prompts that could potentially result in AIG-CSAM.
- The number of instances of CSAM detected at the inputs: 2
- The number of user reports submitted for various model violations: >2,000 models have been reported by Civitai's users
- The number of prevention messages surfaced due to violative prompts: >120,000 (Civitai surfaces prevention messaging every time a prompt is blocked)

- The number of instances of AIG-CSAM detected at the outputs: >100
- The number of reports sent to NCMEC for CSAM and AIG-CSAM: 2, and >100

Invoke

INVOKE REPORTS

According to Invoke, they have four primary strategies to enact this sub-principle, for their 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, these input-level detection defenses are implemented via prompt monitoring, using third party monitoring tools (askvera.io) to ensure they can detect, ban, and report any users attempting to create abusive content on their hosted products. Invoke further reports that whenever violations of acceptable use are detected on their platform, they regularly warn, ban and report users based on the severity of the attempted generation. According to Invoke, they commit 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, they have existing workflows to allow for customer feedback on any and all issues related to the generated media their SaaS solution customers produce using their platform, including any feedback related to content that may contain illegal or unethical material.
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 they make 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, they use this hashlist to ensure that all uploads of these models are automatically blocked.

In regards to their OSS offering, Invoke reports that they have two primary strategies to enact this sub-principle: prevention messaging and model suppression. According to Invoke, both of these strategies are implemented in their OSS offering, in the same way as they are implemented for their SaaS solution. Invoke further notes that for their OSS offering, they have found that this form of open source deployment both allows their business to receive far more QA and testing than may be the norm within their field, but also results in their services being leveraged in ways they 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 their SaaS solution.) In order to meet their 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 their output-level detection defenses of their SaaS solution. In order to meet this commitment, Invoke will need to incorporate detection at the outputs into their SaaS solution moderation system.
3) Customer feedback: Invoke does not currently provide resources for users of their SaaS solution to report violative models they have discovered (e.g. models that can produce AIG-CSAM, or were otherwise optimized for producing AIG-CSAM), to the appropriate organizations. In order to meet this commitment, Invoke will need to provide resources (e.g. guidance on reporting pathways or existing reporting mechanisms) to their customers.
4) Prevention messaging: We currently do not see a gap between what Invoke self-reports as having implemented, and what they committed to implementing.
5) Model suppression: We currently do not see a gap between what Invoke self-reports as having implemented, and what they committed to implementing.

For Invoke's OSS offering, they have not yet incorporated user reporting functionality to allow users of their OSS offering to report violative models they have 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 their OSS offering. Invoke reports they will continue to evaluate ways they can improve their prevention strategy long-term. In order to meet this commitment, Invoke will need to iterate and expand on their user reporting and prevention strategy for their 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: 73.
- The number of user reports submitted for various model violations: 0.
- The number of reports sent to NCMEC for CSAM and AIG-CSAM: 0 and 73.

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

According to Invoke, they do not have any metrics to report for the following items, as they have 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, they have 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 their 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 they build are only used to generate content for the specific use case in which they have 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 they employ human moderators to review every piece of received film studio data for illegal and unethical content. They similarly report employing human moderators to review every piece of generated media for the same purpose.
3) Customer feedback: According to Metaphysic, they have existing workflows to allow for customer feedback on any and all issues related to the generated media they produce for their 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 as having implemented, and what they 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: 0 and 0.

Metaphysic reports that with the above strategies in place, they have not discovered any CSAM or AIG-CSAM produced by their generative AI models, and therefore have 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 they have established terms of service that prohibit customer use and upload of third-party generative AI models hosted on their site to generate AIG-CSAM, other sexually exploitative depictions of children, and photorealistic depictions of children. According to Civitai, they enforce their policies with a combination of human and automated review.

Civitai further reports that when violative models are discovered via user reporting, they are either removed from hosting (see "User reporting" in the "Safeguard our generative AI products and services from abusive content and conduct" sub-principle above) or mitigated once detected. According to Civitai, they leverage semi-permeable membranes [2] (SPMs) for their strategy on mitigation, to ensure that no CSAM or toxic mature content is generated via their cloud-hosted generative AI models. They also report that they restrict certain models to only permit cloud-hosted generation that is subject to filtering and SPM mitigations. Their reported approach involves training multiple SPMs on distinct concepts and then merging these models into one.

More generally, they state that this process of discovering and mitigating third-party models that are capable of generating child safety violative content is currently a very manual process. Civitai reports that there remain challenges to doing this retroactive assessment of models, comprehensively at scale.

NOT YET IMPLEMENTED

Civitai has not yet incorporated mitigations for all of the Stable Diffusion 1.5 models (and its derivatives) hosted on their site. Stable Diffusion 1.5 has been confirmed as able to generate AIG-CSAM [3] and was trained on CSAM [4]. Incorporating mitigations into these models or otherwise removing them from access will be necessary to meet their commitments.

More generally, Civitai is currently leveraging their SPM mitigation only in their cloud-hosted generative AI models, and is not requiring similar mitigations for those third-party models hosted on their site. To meet their commitment on this sub-principle, Civitai will need to ensure that not just their cloud-hosted offerings, but also the third-party models available for download from their site incorporate these same mitigations where necessary.

Further, 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. They are also not

comprehensively retroactively assessing their currently hosted models. Their primary blocker to both of these interventions is a lack of automated model assessment technology. They have done early work ideating on possible solutions (e.g. upon model upload, first put the model in a siloed graphics processing unit (GPU) cluster and assess it with a set of predetermined prompts and automated ML/AI detection at the outputs). They do not currently have the hardware available to do this type of assessment for the scale of models uploaded every day, but report that they hope to have a beta system in place later this year.

Other alternatives for assessment could include making use of existing tags, e.g. as conducted in [5], where the researchers' analysis of Civitai's platform from November - December 2023 found that "Deepfakes" and "NSFW content" are positively correlated for models on Civitai's platform (ϕ = 0.17). 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 their platform.

To meet their commitments, Civitai will need to incorporate systematic model assessment of the third-party generative AI models hosted on their 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 models taken down and removed from platform access, due to discovering they are capable of producing AIG-CSAM and CSEM: Estimated at between 5-10 per month
- Of those models, the number of models for which mitigations were incorporated and the model was re-uploaded: 1

According to Civitai, they do not have any metrics to report for the following item, as they have not yet implemented the underlying interventions:

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

Invoke

INVOKE REPORTS

According to Invoke, they do not serve as a platform for third-party developers to distribute or merchandise their models, nor do they build any first-party models. However, they have 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 their SaaS solution and OSS systems) to generate AIG-CSAM and other sexually exploitative depictions of

children. They further have 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 as having implemented, and what they committed to implementing.

Metaphysic

METAPHYSIC REPORTS
According to Metaphysic, they do not host any third-party models. However, when making use of third-party models internally, they report that they assess every model, prior to using said model, for a variety of ethical issues (including child safety violations). If any issues are found, they do not use the model.

When considering the first-party models they build, as noted in the discussion for the sub-principle "Incorporate feedback loops and iterative stress-testing strategies in our development process," Metaphysic reports they have not yet incorporated red teaming into their processes. However, they do report that they practice model assessment and phased deployment of their models. According to Metaphysic, this model assessment is currently manual. They are 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 their generative AI models.

NOT YET IMPLEMENTED
To meet their commitments, Metaphysic will need to incorporate systematic model assessment of their generative AI models for their capability to produce AIG-CSAM and other child safety violative content.

IMPACT METRICS
Metaphysic reports that 100% of their 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 they have 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 will need to update their model card equivalents to include a child safety section detailing steps the third-party model developer has taken to follow the "Develop" principles. One path they are exploring is requiring third-party model developers to warrant that, before uploading the model, they have curated and cleaned their training data as described in the sub-principle "Responsibly source and safeguard our training datasets from CSAM and CSEM."

Their main concern is the need to balance the right amount of relevance and prominence, such that the information provided does not point bad actors towards the models that were uploaded without safeguards in place. As noted in the discussion for the sub-principle "Responsibly host our models," this could be navigated by using that provided information to determine whether a model is approved for hosting, prior to hosting that model.

Invoke

INVOKE REPORTS

Invoke reports that they do not develop first-party models, nor do they serve as a platform for third-party developers to distribute or merchandise their models, and therefore do not make use of model cards in either capacity. In regards to the third-party models that are uploaded to their SaaS solution or OSS offerings, according to Invoke they offer 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 as having implemented, and what they 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 their customers regarding what child safety interventions were put into place as part of model development.

Metaphysic

METAPHYSIC REPORTS

According to Metaphysic, they have incorporated into their 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 as having implemented, and what they committed to implementing.

IMPACT METRICS

Metaphysic reports that 100% of their 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. Further, Civitai's reports to NCMEC include the model used to generate the offending image, when that information is known. Additionally, Civitai reports they have updated their policies such that any resource or workflow advertising itself as nudifying "real images" of "real people" is disallowed. According to Civitai, they enforce these policies using the same strategies described in the sub-principle "Safeguard our generative AI products and services from abusive content and conduct." Additionally, Civitai reports that their existing policies against "suggestive" or "sexual" content depicting real people, combined with their use of prompt filters and SPMs for cloud-generated images captures a significant portion of nudifying activity.

NOT YET IMPLEMENTED

Civitai does not retroactively check their existing corpus of hosted models, to confirm that the newly discovered problematic models do not appear anywhere else in their collection (as opposed to their current suppression strategy that helps ensure the model is never added back into their collection). To meet their commitment on this sub-principle, they will need to add in these retroactive checks as well.

In regards to their policy and enforcement mechanisms around "nudifying" services, regardless of how these nudifying services are advertised (e.g. for use on children, for use on adults, for use on real people vs. fictional characters), the underlying technologies can and are being used to nudify children [7]. To meet their commitment on this sub-principle, Civitai will need to update their policies and enforcement mechanisms to either explicitly disallow "nudifying" models and services more

holistically, or otherwise devise a strategy such that those nudifying services and models hosted on their site are not misused for nudifying depictions of children.

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: Estimated at between 5-10 per month
- The number of prevented attempts to upload a model optimized to produce AIG-CSAM: >400

According to Civitai, they do not have any metrics to report for the following items, as they have not yet implemented the underlying interventions:

- The number of nudifying services, retroactively removed
- The number of prevented attempts to upload a nudifying model or nudifying workflow

Invoke

INVOKE REPORTS

According to Invoke, they make 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, they use this hashlist to ensure that all uploads of these models (in both their SaaS solution and their 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 they retroactively check uploaded models in their 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 they do 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 their strategy for addressing this type of misuse.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Invoke self-reports as having implemented, and what they 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 their SaaS solution offering, as Invoke does not have telemetry or access to collect metrics for their OSS platform. Invoke further notes that they have never had a user attempt to upload to their 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, they do not have any metrics to report for the following items, as they do 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, they do not host third-party models or services, or offer search functionality as part of their business model.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Metaphysic self-reports as having implemented, and what they 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, they have invested in and deployed future technology solutions via their line of work around SPM-based interventions. They further report that they monitor their user community for emerging risks, and rely on outside partners to also monitor trends and emerging risks. Additionally, they report continuous effort improving the ML/AI detection technology they build in-house.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Civitai self-reports as having implemented, and what they 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.
- Cadence at which mitigations are assessed against the business' tech stack, to ensure effective performance: Weekly or daily, depending on the scale of reports and automation needs.

Invoke

INVOKE REPORTS

According to Invoke, they have invested in and deployed future technology solutions via their work on Vera (askvera.io), a tool to perform prompt monitoring and detection. Invoke further notes that as this is a set of tools they use everyday, they regularly maintain and update these tools.

In addition, Invoke reports that they leverage their 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 their 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 as having implemented, and what they 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: $100,000 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 their generative AI models. As a result of this controlled access, Metaphysic reports they have not made use of open source intelligence (OSINT) or other strategies to understand how bad actors are potentially misusing their products and services. In regards to investing in research and technology, Metaphysic reports that they intend to publish their findings around their efforts to build ML/AI dataset segmentation technologies. They further report (as outlined in the discussion on "Responsibly host our models") their investment in building scalable, automated model assessment mechanisms.

NOT YET IMPLEMENTED

Currently, we do not see a gap between what Metaphysic self-reports as having implemented, and what they 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, the same four strategies outlined in "Safeguard our generative AI products and services from abusive content and conduct," are employed generally across their site. According to

Civitai, when reporting AIG-CSAM to NCMEC their content moderation team ensures that their 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," Civitai is not yet using hashing and matching against third-party owned, maintained and verified CSAM lists to detect known CSAM hosted on their platform. Additionally, Civitai does not yet employ prevention messaging as part of their safeguarding the search functionality on their site (e.g. entering the terms "child abuse model" into their in-site search functionality does not surface prevention messaging). To meet their commitment, Civitai will need to incorporate hashing and matching against verified CSAM lists in their overall content moderation strategy, as well as incorporate prevention messaging for the search functionality on their site.

IMPACT METRICS

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

- The number of instances of CSAM detected on their site: 2
- The number of user reports submitted for various violations on their site: >200,000
- The number of instances of AIG-CSAM detected on their site: >100
- The number of reports sent to NCMEC for CSAM and AIG-CSAM as a result of the above: 2 and >100

According to Civitai, they do not have any metrics to report for the following items, as they have 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 their SaaS solution and OSS offerings. According to Invoke, when reporting AIG-CSAM to NCMEC their content moderation team ensures that their 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 their 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 their commitment, Invoke will need to incorporate CSAM detection in their overall content moderation strategy.

IMPACT METRICS

According to Invoke, they do not have any metrics to report for the following items, as they have 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, they do 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: 3-Month Progress Report on Civitai and Metaphysic." *Thorn*, 26 Sept. 2024, https://www.thorn.org/blog/safety-by-design-for-generative-ai-3-month-progress-report.
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. Portnoff, et al. (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.