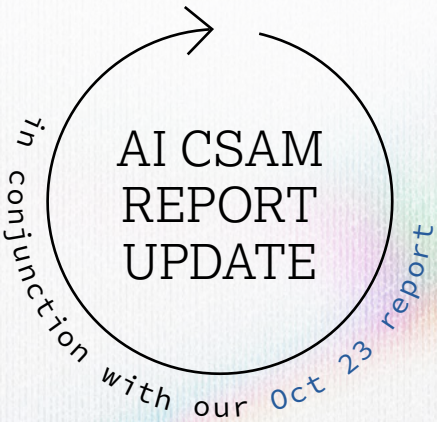
IWF
Internet
Watch
Foundation

July 2024 | V1.1

# What has changed in the AI CSAM landscape?




**Prompt:** from fantasy to photo-realistic reality

PUBLIC VERSION

# Content note

Throughout this report update, Child Sexual Abuse Material generated through Artificial Intelligence is referred to as **AI CSAM.**

This report update contains no AI CSAM.

It contains descriptions of the methods used to generate AI CSAM, alongside other verbatim comments from perpetrators.

The verbatim comments from perpetrators are reproduced in the report update as they were typed on screen.

# Table of Contents

IWF

Please click on the IWF logo **'home' button** at the top of each page to navigate back to the contents page.

INTERACTIVE REPORT UPDATE

# Foreword from Susie Hargreaves OBE

**Olivia is a little girl who we have seen grow up online through images uploaded of her abuse at the hands of a sex offender when she was between three and eight-years-old. We told her story in 2018 in our annual report. The updated story I'd like to tell you now, is that after Olivia was rescued all images showing her sexual abuse were deleted. But that's not the case.**

Yes, she was rescued from her abuser, but now new images of Olivia are being created by individuals who want to see her in new, abusive, situations, years after the physical abuse ended. And the AI tools to make this happen are here, now.

One of our top priorities for a future Government is to get a grip on the impending child sexual abuse crisis online exacerbated by new technologies.

This report sets out the progress made since we last reported on the impact of generative AI in October 2023, and powerfully makes the case for why this must be a top priority.

I am delighted the Online Safety Act finally made it onto the statute books in 2023, hot on the heels of the Digital Services Act in Europe. We have not seen enough progress, however, to deal with the impacts of generative AI technology. Nor have we seen sufficient clarity that these new technologies will be developed with safety in mind or effectively regulated.

It has been almost a year since we sounded the alarm, and this updated report identifies new challenges with generative AI. Technology companies unabashed march to text-to-video creation; the announcements by OpenAI that they are consulting on the possibility of their tools being used to create Not-Safe-For-Work content, and the emergence of offenders using LoRA models of known victims of child sexual abuse are just some of the most pressing challenges we face.

In April 2024, the Government committed to criminalising the creation of deepfakes. Of course, in the child safety space, deepfake child sexual abuse images are already illegal, but the technology used to nudify children isn't.

We welcomed this news as it saw two of the largest nudifying apps, which have amassed millions of users, and have been used to create child sexual abuse material, being disabled to public access in the UK. Less welcome, was the fact this is still yet to become law because of the timing of the General Election.

Two other amendments which were due to make it into law as part of the Criminal Justice Bill also fell: one sought to extend the existing offence of possession of a paedophile manual to cover the exchanging of hints and tips on the abuse of generative AI tools. The other proposed tackling AI chatbots which seek to simulate the offence of sexual activity with a child. These were both widely supported in Parliament.

The Data Protection and Digital Information Bill had an amendment tabled to it which sought to make it an offence to train an AI tool on child sexual abuse material or for an AI tool to generate child sexual abuse material. But this also failed to become law.

Whilst it was disappointing to see these legislative developments not make it on to the statute books, we are pleased to see that there is cross-party consensus on tackling these issues. The Labour Party has said it will: "ensure the safe development of AI models, by introducing binding regulation on the handful of companies developing the most powerful AI models and by banning the creation of sexually explicit deepfakes" in its manifesto. The Conservatives also committed at Report Stage during the Criminal Justice Bill to the intention to bring back amendments at a later stage.

This report highlights how desperately the law needs to change to keep pace with technology. It also provides an update on previous recommendations, continues to chart the impact of generative AI on the spread of child sexual abuse and makes further recommendations for an incoming Government.

We will be watching closely to see how industry, regulators and Government respond to the threat, to ensure that the suffering of Olivia, and children like her, is not exacerbated, re-imagined and re-created using AI tools.

**Susie Hargreaves OBE** | IWF CEO

# New (additional) recommendations:

**FOR GOVERNMENT:**

1. That the Government legislates to ensure that paedophile manuals which exchange hints and tips on how to utilise text-to-image based generative AI tools to create child sexual abuse material are made illegal, by extending the existing offence, to cover pseudo images.
2. That the Government legislates to make it an offence to use personal data or digital information to create digital models or files that facilitate the creation of AI or computer-generated child sexual abuse material.
3. That the Government legislates to tackle the rise in generative AI chatbots which simulate the offence of sexual communication with a child.
4. That the Government legislates to ensure nudifying technology is not available to UK based users and encourages other Governments globally to take similar measures.

**Relevant passages which relate to the above recommendations are highlighted throughout this report.**

# Update on previous recommendations:

**FOR GOVERNMENT**

| Previous Recommendation | Progress to date |
|---|---|
| **Explore at the forthcoming AI summit the challenges for dealing with AI CSAM, including the need for international alignment.** | • IWF and Home Office, including then Home Secretary, Suella Braverman, jointly hosted an AI Safety Summit fringe event at Chatham House in London, two days before the AI Safety Summit at Bletchley Park.<br>• 33 NGOs, tech companies, Governments, law enforcement and academics agreed a non-binding pledge to tackle AI generated CSAM at the summit.<br>• G7 communique highlighted the challenges of Artificial Intelligence and a commitment to working together to align internationally.<br>• IWF has presented this research in Japan and the US.<br>• Law change announced in Europe through changes to the Directive, laying down new rules to tackle Child Sexual Abuse. |

| | |
|---|---|
| **Ministry of Justice review of laws ensuring they are fit for the AI age** | • Three amendments tabled to the Criminal Justice Bill to tackle the impacts of AI generated CSAM and nudifying technologies.<br>• One amendment tabled to the Data Protection and Digital Information Bill.<br>• No substantive law changes yet, but commitments from both Labour and the Conservatives to address these issues if they form the next Government.<br>• Labour frontbench spoke in support of this recommendation at Report Stage in the House of Commons on the Criminal Justice Bill. |
| **To consider extension of IWF remit to be able to scrutinise datasets on which these technologies are trained.** | • Crown Prosecution Service has confirmed that IWF has the relevant authority to process and scrutinise AI models.<br>• Work underway with Government departments to see what further support IWF can give. |

**FOR LAW ENFORCEMENT AND REGULATORS**

| Previous Recommendation | Progress to date |
|---|---|
| **For the College of Policing training course to be updated to cover AI CSAM and ensure clear guidance is issued to police graders.** | • The College of Policing is currently liaising with the Crown Prosecution Service before introducing additional guidance so officers can more effectively grade child sexual abuse images in accordance with national guidelines.<br>• The College of Policing also provides a standalone learning product on deep fakes which was released to forces in early 2024. |
| **To ensure there is proper regulatory oversight of AI models before they go to market and ensure mitigations are in place for open-source models with closed source having protections built in.** | • In August 2023, the UK Government published a white paper entitled: **"A pro innovation approach to AI regulation."** They published their response to this White Paper in February 2024, with the Government concluding: "It will not rush to legislate or implement 'quick fix' rules that would soon become outdated or ineffective. Instead, the government's context-based approach means existing regulators are empowered to address AI risks in a targeted way."<br>• Looking ahead to manifesto commitments made by political parties at the 2024 UK General Election on Artificial Intelligence:<br>**The Labour Party** has said: "ensure the safe development of AI models, by introducing binding regulation on the handful of companies developing the most powerful AI models and by banning the creation of sexually explicit deepfakes." |

| | |
|---|---|
| | **The Conservative Party** has said: "The UK is well positioned to spearhead this transformation and is already leading global work on AI safety. Over the last 14 years, the Conservatives have turned the UK into a science and innovation superpower."<br><br>Along with a commitment to: "Building on existing responsibilities set out for social media in the Online Safety Act."<br><br>**The Liberal Democrat Party** has said they will: "Create a clear, workable and well-resourced cross-sectoral regulatory framework for Artificial Intelligence that:<br>• Promotes innovation while creating certainty for AI users, developers and investors.<br>• Establishes transparency and accountability for AI systems in the public sector.<br>• Ensures the use of personal data and AI is unbiased, transparent and accurate, and respects the privacy of innocent people."<br>• In Europe, we have seen the European Institutions pass into law the first piece of legislation to regulate Artificial Intelligence. |

**FOR TECH COMPANIES**

| Previous Recommendation | Progress to date |
|---|---|
| **To ensure that companies using and developing Generative AI and Large Language Models (LLMs), place clearly in their terms and conditions that the use of these technologies to generate child sexual abuse material is prohibited.** | • Stability AI, OpenAI and many other platform's Terms and Conditions have been clear that the use of their technologies to create child sexual abuse material is prohibited.<br>• We have seen OpenAI announce a consultation into the possibility of its technologies being used to create content that is not safe for work.<br>• LAION, one of the biggest providers of open-source data sets has established a relationship with the Internet Watch Foundation and other child safety organisations.<br>• Stability AI has become the first member from the Artificial Intelligence sector to join the IWF as a Member. |
| **That search services should de-index links to fine-tuned AI models known to be linked to the creation of AI CSAM.** | • Thorn, All Tech is Human and the major developers of AI technology have all committed to a set of voluntary principles to make AI safe by its design. |

# Executive summary

Use of Artificial Intelligence (AI) to generate child sexual abuse material (CSAM) is increasing, and the technology is fast improving.

The dark web child sexual abuse forum surveyed in October 2023 was revisited, and a new analysis found that:

- More criminal AI CSAM images were shared – a total of **3,512 AI CSAM images.**
- **90% of images assessed by IWF analysts were realistic enough to be assessed under the same law as real CSAM.**
- Those images **contained more images in the most severe category of CSAM in the UK** (Category A, which contains penetrative sexual activity, bestiality, or sadism) than in October 2023 – this time, 32% of criminal pseudo-photographs were Category A, indicating that perpetrators are experiencing more success generating complex 'hardcore' scenarios.

**Other findings:**

- **The first AI CSAM videos are now in circulation.** These are mostly partially-synthetic – 'deepfake' – videos, though some primitive fully-synthetic videos also exist.
- The IWF has been encountering **an increasing amount of AI-generated content, including AI CSAM, on the clear web.**
- Extensive evidence for **the sharing of AI models for generating images of specific children,** including known victims of CSAM and famous children, has been identified, and is provided in this report update.

3

# Introduction to this report update

In the summer of last year (2023), the Internet Watch Foundation (IWF) first reported that open-source AI models were being widely used to generate CSAM.

A report was compiled and was released in October 2023. It found that perpetrators were able to download – legally – everything needed to generate lifelike images of child sexual abuse, then produce as many of those images as they desired. Generation of AI CSAM took place offline, with no opportunity for detection.

The report found evidence of the sharing of thousands of those images, particularly on the dark web – images that comprised new threats both towards existing victims of child sexual abuse and towards potential new victims of child sexual abuse.

This report update seeks to describe what has changed in the AI CSAM landscape since then. It should be considered an update to the October 2023 report, to be read in conjunction with it.

Since autumn last year, some progress towards highlighting and prioritising child safety in AI development has been made. Collaborative efforts among government, law enforcement, the technology industry and civil society have forged valuable channels of communication, and have begun a process towards recognising that AI left unchecked has the potential to corrode child protection efforts. The first steps have been taken towards urgently-needed preventative and mitigative action.

As with all online safety challenges, this challenge is inherently international. It is encouraging that the UK government has sought to position the country at the forefront of AI safety and regulation in hosting the first international conference on the issue, the AI Safety Summit, last November. The Republic of Korea hosted the 2024 AI Seoul Summit in May.

This report update shows that the pace of AI development has not been slowing, nor has the number of people using AI for criminal purposes decreased. In this context, and in the context of the better, faster, and more accessible tools to generate images and videos, the future continues to hang in the balance.

AI still poses a significant risk to the IWF's mission to remove child sexual abuse material from the internet. It still has the potential to overwhelm

resources and cause irreparable harm to children. But the right decisions made now – to necessitate safety by design, to ensure rigorous testing of all AI models released to the public, and to put protection of children before pursuit of profit – can mitigate these problems for years to come.

## Notes on terminology

As in the October 2023 report, this update uses the term 'AI CSAM' to refer to criminal images or videos of the sexual abuse of children that are generated or edited by AI technology, and 'real CSAM' to clearly distinguish CSAM that is not generated or edited by AI technology.

The term 'deepfake' is used variously in the AI field, in the media, and among the wider population. Sometimes it is taken to refer to all AI-generated or AI-edited content. This report uses the term 'deepfake' to refer to *partially-synthetic* content: edited content that is based on a real image or video but has been altered using AI technology. This is particularly important in the context of 'deepfake videos' – in this report update, edited (or 'faked') real videos – which should be clearly distinguished from fully-synthetic videos created by text-to-video or text-to-image-to-video.

## Outline and guide for readers

Section 4 of this report update tracks shifts in the use of newer, higher-quality versions of open-source image-generating tools; notes the progression of AI-generated video content, including AI CSAM video; and details evidence for the spread of AI CSAM across the clear web.

The October 2023 report included a study of a dark web CSAM forum, in which all the AI-generated images posted to the forum in a one-month period were scraped and assessed against UK law. In section 5, this forum is revisited, and a new scrape is completed. This allows for a comparative analysis, which asks whether there have been changes in the type or quality of the imagery shared.

**Newly for this report update, available metadata is scraped from these images.** This metadata is analysed to assess how far the methods used to generate those images can be deduced.

The IWF extends its thanks to Camera Forensics for their assistance on image metadata analysis.

**For further information on the IWF and its remit, see the October 2023 report.**

4

# Trends monitoring

## IWF reports and the clear web

The IWF continues to track reports (generally, webpages) containing AI-generated content, including AI CSAM. Most of these reports are received from members of the public; a small number arise from analysts' proactive searching for content. Most reports from the public continue to contain non-criminal content only.

The graph below shows a general gradual rise in the number of reports where any kind of AI-generated content has been identified by an analyst (in red) alongside a subset of these figures, reports where AI CSAM has been identified and 'actioned' – processed as criminal under UK law (in grey). The IWF seeks to get 'actioned' AI CSAM removed from the internet.

**IWF reports containing AI-generated content, March 2023 to April 2024**

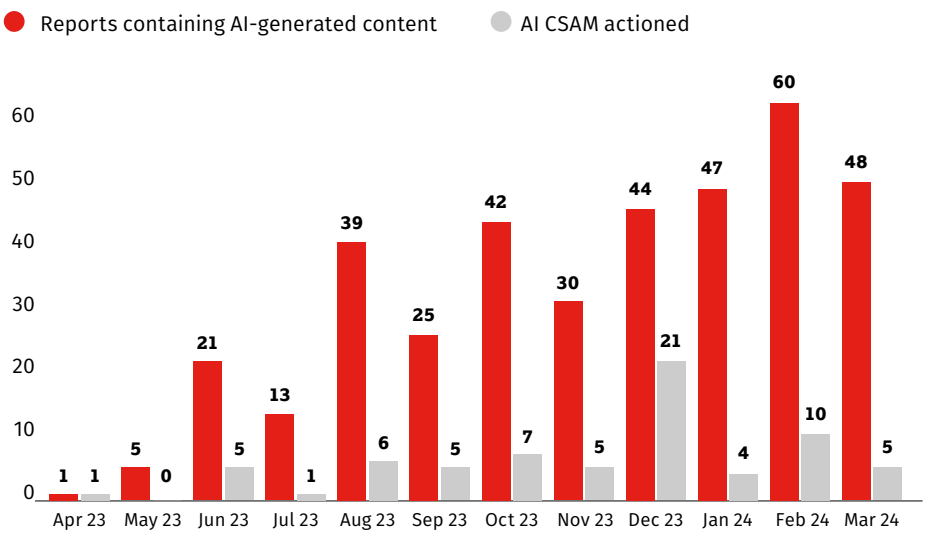



**Figure 1**
Reports to IWF containing AI-generated content rose gradually from April 2023 to March 2024.

Source: IWF analysis

**Reports from April 2023 to March 2024 total 375 reports that contained AI-generated content (and peaked at a high of 60 reports in February of this year) - of which 70 reports contained criminal AI CSAM. These reports almost exclusively point to content hosted on the clear web.** They provide some evidence for the spread of AI-generated content and AI CSAM across the publicly accessible internet.

Notably, we have seen AI CSAM images being shared on commercial sites on the clear web – in place of 'real' images of children. These include on dedicated commercial banner sites, and forum pages with affiliate file-hosting links.

**We also actioned the first UK-hosted webpage containing AI CSAM in March 2024.**

Reports containing AI-generated content continue to comprise a low proportion of total IWF reports. In 2023, this was 220 out of 392,660 reports (0.06%). Of the public reports – likely a better metric to indicate the prevalence of AI-generated content online, or the likelihood of individuals encountering AI-generated content online – this was 201 out of 123,667 (0.16%).

Statistics from the US-based National Centre for Missing and Exploited Children (NCMEC) for 2023 paint a similar picture[1]: generative AI featured in just 4,700 out of over 36,000,000 reports (0.01%). (Note that the comparison is not isomorphic as NCMEC reports are collected and processed in different ways; nonetheless, the low proportion is notable.)

It is fair to state that the rise of AI-generated content over the past year has been gradual and not exponential in nature, and that generative AI – though fast establishing itself within this space – has not yet broken into the 'mainstream' either of adult pornography or of CSAM (insofar as a mainstream exists).

**The October 2023 report highlighted two notable categories of user-friendly online AI pornography services, both of which feature heavily among public reports to IWF: (1) fully-synthetic pornography generation services, and (2) 'nudifying' services.**

SEE RECOMMENDATION #3 & #4

For Government

Another important category of public reports received related to AI chatbots, which tend to feature a variety of 'characters' with which to interact – often including underage characters. Use of chatbots to simulate conversation with a child is somewhat outside IWF remit, though anecdotal evidence of perpetrators exchanging models, tips and advice has been found.

Use of chatbots in this way has the potential to encourage or normalise harmful behaviour among those with a sexual interest in children. Many of these chatbots are accessible with no age verification process; others incorporate fully customisable (or open-source) characters and have few or no limits to the topics or content of conversation. Evidence of effective regulation in this space is negligible.

### 1. AI pornography services

Some barriers to entry to generating images with AI text-to-image models persist: the ability to learn the required technical skills, alongside the possession of some level of computer hardware. (These two important barriers may go some way to explaining the limited spread of AI CSAM so far observed.)

Services for AI pornography aim to neutralise these barriers by making generation fast and intuitive. Users do not have to download or run programs; they simply type or select what they want to see, and the service

1. www.missingkids.org/content/dam/missingkids/pdfs/2023-CyberTipline-Report.pdf

– which usually uses a built-in foundation AI model version as the image generation 'engine' – provides the images. Evidence of (presumably, low-tech) perpetrators trying and failing to generate AI CSAM on these platforms has been found shared on dark web forums and was included in the October 2023 report. Nonetheless, two 'actioned' reports between April 2023 and March 2024 – webpages containing criminal AI CSAM – relate to two of these AI pornography sites, showing that it is possible for perpetrators to succeed in abusing these services to generate AI CSAM.

**2. 'Nudifying' services**

SEE RECOMMENDATION #4

For Government

**The prevalence of 'nudifying' platforms has been increasing over the past year – indicated by rising public reports to IWF, but also simply by the volume of these sites on the internet. In short, these are sites in which a user uploads an image of a clothed individual; the model outputs an interpretation of the individual without clothes. These comprise one category of 'deepfake' sexually explicit images, soon to become a criminal offence in the UK.**

**Between April 2023 and March 2024, 21 public reports specified dedicated 'nudifying' websites across 15 different domains.**

At the same time, an increasing number of those whom IWF terms 'self-reporters' – members of the public reporting their own explicit imagery – have reported 'fake' content of themselves. **In 2023, the IWF received 17 self-reports from members of the public referencing 'fake' or 'AI-edited' content of themselves. Many of these are children.**

Anecdotal evidence suggests that the perpetrators in these cases are generally people who are unknown to the reporters – whose relationships to the reporters are online-only – but the data is too limited to draw firm conclusions on this point.

Among these cases, the IWF has seen evidence of 'innocent' (non-explicit) imagery of children being taken and 'nudified'.

Sometimes, these 'nudified' images are posted on social media sites with the intention of being shared and seen widely, to cause the victim more distress.

Reports of fakes and deepfakes – many of which are generated using these 'nudifying' services – seem to be closely linked with reports of financial 'sextortion', or blackmail with sexually explicit images. The crux of this point is that perpetrators no longer need to source intimate images from children because images that are convincing enough to be harmful – maybe even as harmful as real images in some cases – can be produced using generative AI.

SEE RECOMMENDATION #1

For Government

**Indeed, one 'paedophile guide' identified by IWF contained a section explicitly encouraging perpetrators to use 'nudifying' tools to generate material to blackmail children. The author of this guide claimed to have successfully blackmailed 13-year-old girls into sending intimate images.**

# AI videos

### Fully-synthetic videos

Recent months have seen notable progress towards fully-synthetic realistic video content in the form of new video generation models, including Stable Video Diffusion (November 2023), a preview of Sora (February 2024) and a preview of Veo (May 2024).

OpenAI's Sora can generate convincing minute-long videos from text, image, or video. It has been released to limited researchers, with a public release planned this year. This type of release, a research-only preview in advance of a full public release, has also been employed for Google's new state-of-the-art video generation model, Veo. These two – crucially, closed-source – models sit at the current frontiers of the video generation industry.

**Figure 2**
A screenshot from a video generated by OpenAI's Sora shows a woolly mammoth walking in the snow.

Source: OpenAI

Stable Video Diffusion can generate short (seconds-long) videos from images. As a Stability AI product, it has been released as an open-source model – available to all under a non-commercial licence. Output is comparable to the closed-source RunwayML Gen-2.

What is the ultimate goal for video-generation companies like RunwayML?

> ***"We've always set the ability to generate a two-hour film as a north star."***

### Text-to-video CSAM

Perpetrators watch the latest advancements with interest. On a dark web forum, AI CSAM perpetrators discuss AI-generated videos:

> *"How long until we can use this new Sora software to make whatever video we want? I want to put my sister's photos in from when she was a kid and make her do nasty things"*

> *"Am seeing the video trailers that were generated by AI, and my mind is blown... The ability to create any child porn we desire... our wildest fantasies... in high definition."*

Limited moving image (GIF) and video CSAM has so far been seen but has been slowly increasing over the past months. Some can be described as deepfakes – for the purposes of this report update, partially-synthetic content – which are discussed later in this section. Fully-synthetic AI CSAM videos are rarer, and are fairly primitive.

One 18-second video, found shared alongside almost 5,000 AI-generated images, shows an adult male penetrating a girl, approximately 10 years old. She is sitting on top of him and looking at the virtual 'camera'. Behind them is a well-lit room with large windows.

The video flickers and glitches; her face and expression morphs from frame-to-frame. Her movement is jerky. Nonetheless, the activity is clear and continuous. It is obvious that the video is synthetic – it doesn't look much like a real video – but this was also the case with images two years ago.

These observations mirror the wider state of AI-generated adult pornography videos: convincing deepfake videos, and primitive (but fast-developing) fully-synthetic videos.

## Technology, tools and models

**Open-source models**

Perpetrators continue to use open-source models as the tool of choice to generate CSAM images because – as described in the October 2023 report – access is offline, on-device; users can use models and prompts freely; and there are few (if any) opportunities for content moderation and criminal content detection or prevention.

There is, however, some evidence that perpetrators are moving away from earlier versions and using more recent image generation models in increasing numbers.

Anecdotal evidence has been found of some perpetrators sharing AI-generated images of children – including AI CSAM – and claiming to have used more recent foundation models only (in other words, no fine-tuned models at all).

An increasing number of fine-tuned models shared in AI CSAM communities are also intended for use with the latest foundation models, resulting in images being generated using these versions. As one user asks another sharing an older model.

> *"Have you considered making a model with [redacted]? In my experience it is generally a big improvement over [redacted]."*

Another user comments:

> *"[redacted] based checkpoints are already at the point where there are some pictures that I wouldn't be able to identify as being AI."*

In the dark web CSAM forum from which images were scraped for this year's study, 35% of (apparently CSAM-trained) models whose links were directly shared were for [redacted]; the remainder were based on older versions. It is, however, notable that it is possible to use those older fine-tuned models with [redacted].

Such evidence of misuse is despite open-source AI models working to implement multiple safety features for their foundation models, including filters for 'unsafe' content. Indeed, AI CSAM perpetrators on dark web forums are dismissive about purported safety features in newer models:

> *"If there is any alignment training inherent to [redacted], large finetunes will be able to override it."*

### Deepfake videos

As set out in the introduction, this report update takes deepfake videos to be partially-synthetic videos – generally, videos *edited* using AI tools to add the face or likeness of another person. It remains the case that the overwhelming proportion of deepfake videos are pornographic in nature. The best deepfake videos are now almost seamless – containing little visual evidence of modification.

A large amount of media discussion concerns abuses of celebrity likenesses in deepfake videos – indeed, these comprise most of the videos on the largest deepfake pornography websites, some of which attract millions of visitors every month. Some of the most-publicised cases involve abuses of celebrity likenesses in pornography; others concern use for misinformation, or for humour or entertainment.

**Figure 3**
'DeepTomCruise' before/after comparison shows the application of deepfake technology to viral video content from 2021.

Source: The Verge

Nonetheless, this same technology can be – and is – applied to less well-known individuals, including non-celebrities, and including children.

**Some deepfake CSAM videos shared in dark web forums take an adult pornography video and add a child's face. Others take existing CSAM videos and add a different child's face to them.** Because the original videos are of real children, and have, therefore, real child anatomical proportions, they can be especially convincing. One impressed forum user says:

> *"I knew about deepfakes... This is so on point! The colours, the shadings, no glitch. Truly mind blowing."*

From anecdotal evidence gathered, methods used to generate these videos appear to be the same as those used to generate deepfake adult pornography. One perpetrator claims:

> *"I'm using [redacted] . Go to [redacted] and look in the [redacted]. Everything you need to do this is there."*

Free, open-source AI software is behind many viral deepfake videos and faces the same inherent challenges as open-source foundation models over malicious or illegal use, including for non-consensual pornography (the overwhelming majority of existing deepfakes) and CSAM.

## Dark web forums and AI CSAM discussion

Given the gradual increase in the number of AI CSAM reports on the clear web, dark web forums remain the main hub for IWF for intelligence-gathering on many aspects of AI CSAM.

Dark web CSAM forums are mostly concerned with the sharing and discussion of real CSAM. AI CSAM remains a small – but likely growing – part of these more general forums.

There remain large variations in the level of interest in AI CSAM among these wider CSAM communities. In a recent exchange, an AI-generated image was posted in a section intended for real images, and was met with a mixture of apathy and antipathy:

> *"Nice but it looks like AI-gen and none of us want that."*

> *"100%, only want the real stuff."*

> *"Thanks for these but I'm not into AI."*

SEE RECOMMENDATION #1

For Government

**A common thread across various AI CSAM communities relates to requests for guidance or training, as briefly set out in the October 2023 report. Where people new to AI encounter AI CSAM, they are sometimes impressed:**

> *"They look very real, like you've taken photos of them."*

Perpetrators encourage people towards trying certain generative AI models:

> *"If you are undressing little girls, I think your only mostly safe option is running [redacted] locally."*

> *"Generating on-topic [CSAM] content is the same as any other content, just with a different prompt. After that it's just a lot of experimentation!"*

AI text-to-image models, though, can be daunting for those starting out. Those people, then, ask for advice, tutorials or guides:

> *"Just wondering if you have any tutorials, or how someone can get started making their own pictures?"*

> *"I am very interested in learning how to use AI to develop child porn... just show me the step by step"*

> *"Wanted to know if someone could point me in the right direction to learn, download the software, etc."*

SEE RECOMMENDATION #1

For Government

**At time of writing, the UK prohibition on paedophile manuals continues to exclude pseudo-photographs of children – necessarily encompassing all AI CSAM. This means, therefore, that tutorials and guides shared among members of these communities detailing how to generate realistic AI CSAM remain legal.**

## AI CSAM featuring known victims and famous children

As discussed in the October 2023 report, AI CSAM perpetrators regularly use AI models to generate images of existing children – and the majority of CSAM fine-tuned AI models are designed for generating their images. These are usually known victims of child sexual abuse or famous children.

Perpetrators on dark web forums continue to discuss how to train LoRAs (fine-tuned models) for those named victims or celebrities, share models they have trained and images they have generated, and request new ones.

AI CSAM images of celebrity children may have a broader appeal than images featuring known victims of CSAM. Such images – including some 'packs' of AI CSAM celebrity images – have been seen multiple times on sites on the clear web. Variously, they feature famous children and de-aged famous adults.

It is possible that the world of fine-tuned CSAM models – including those for generating images of named children – runs much deeper than is apparent from looking only in publicly accessible areas on the clear web and the dark web. This is a world that reaches into homes with non-internet connected devices, and – crucially – into end-to-end encrypted, peer-to-peer networks that are inaccessible to organisations like the IWF.

SEE RECOMMENDATION #2

For Government

**One user, seemingly mostly active in these peer-to-peer networks, shared an anonymous webpage containing links to fine-tuned models for 128 different named victims of child sexual abuse.**

Every 'child model' had [redacted] variations; many also had options for younger or older versions of the child in question.

One of these 'child models', 'Olivia', was featured by IWF back in our 2018 Annual Report. In that report, an analyst recounted:

---

***"I first saw Olivia when she was about three.***

***I've seen Olivia grow up through cruel images and videos, suffering hideous abuse. She was repeatedly raped and sexually tortured.***

***We see Olivia every day—five years after she was rescued. To show exactly what 'repeat victimisation' means, we counted the number of times we saw Olivia's image online during a three-month period. We saw her at least 347 times. On average, that's five times each and every working day."***

An AI model for generating novel images of Olivia is available to download for free, just a couple of clicks away. The user can choose to use a particular version of the model. It's a potentially 'popular' model among AI CSAM communities – as one user asks elsewhere,

> *"Anyone trained a LoRA for [Olivia] yet? Would be really cool to see"*

That user is pointed towards the anonymous link that IWF has identified.

Before the advent of AI-generated images, survivors of childhood sexual abuse like Olivia already had to contend with the potential for the images and videos displaying their abuse being shared across the internet. Each time one of those images was shared or viewed added another link in a long chain of child sexual abuse.

Fine-tuned models like Olivia's have been trained on the imagery that IWF was seeing five times a day in 2018 but was unable to eradicate. The consequence of this is a new way of adding links to the chain – each time re-victimising survivors of child sexual abuse – and potentially without end, since perpetrators can generate as many images of those children as they like without fear of detection or prevention.

As explained in the October 2023 report, these are lifelike images – they look like images of real-world abuse – but can produce 'unreal', unseen settings, scenarios, and sexual activities.

SEE RECOMMENDATION #2

For Government

These models fine-tuned on CSAM victims – including Olivia's – remain legal in the UK.

# AI CSAM image analysis: new snapshot study

## Overall forum trends

Part of the October 2023 report comprised a snapshot study of a dark web CSAM forum. For that report, all the live AI-generated images posted to the forum in a 30-day period (September 2023) were identified, and a selection were assessed.

This update revisited the same dark web CSAM forum to analyse whether use of the forum had changed in type or frequency; whether any trends in imagery could be identified; and whether discussions among forum users had progressed.

This new snapshot took the live AI-generated images posted to the forum over another 30-day period (9 March to 7 April 2024) – this time, all the images that were found were assessed by IWF analysts.

The table below compares findings on posts of AI-generated imagery to the forum, and on AI-specific threads (sections where users post content) within the forum, over the two periods.

| | September 2023 | March-April 2024 |
|---|---|---|
| AI-generated images posted (incl. duplicates) | 20,254 | 13,906 |
| AI-generated videos posted | 0 | 9 |
| Count of threads to which (live) AI-generated content was posted | 74 | 106 |
| Sum of views on AI-specific threads created over period | 261,920 | 319,141 |

These findings show that the number of AI-generated images posted to the forum decreased from September 2023 to March 2024. These were distributed across more threads. Images were, then, generally shared in smaller, more 'curated' sets.

The number of views on AI-specific threads created over the two periods increased by 22%. If view count is some guide to general interest, it may be concluded that the level of interest in AI CSAM has increased slightly among users of this forum. (Nonetheless, data on number of unique users is unavailable, so it is impossible to say whether this shows that more people are interested in AI CSAM.)

**9 AI-generated deepfake videos were found to have been posted within the period analysed for this snapshot.** These are not the first AI-generated videos found by IWF on dark web forums, but it is notable that none were found shared here six months previously.

## Image analysis

For this new snapshot study, 13,906 online images were identified and downloaded. After de-duplication, **these totalled 12,148 unique AI-generated images.**

12 IWF analysts dedicated a combined total of 130.5 hours to assessing these 12,148 images.

As outlined in the October 2023 report, AI CSAM in the UK falls under two different laws, which have different criteria and sentencing guidelines:

- The **Protection of Children Act 1978** (as amended by the Criminal Justice and Public Order Act 1994). This law criminalises the taking, distribution and possession of an "indecent photograph or pseudo-photograph of a child".
- The **Coroners and Justice Act 2009.** This law criminalises the possession of "a prohibited image of a child". These are non-photographic – generally cartoons, drawings, animations or similar.

The key criterion for classification as criminal under the Protection of Children Act 1978 is that the image "appears to be a photograph".

**2,985 images were classified as indecent pseudo-photographs, and 527 images were classified as prohibited images – in total, this is 3,512 AI CSAM images.**

These are shown as proportions of the 12,148 assessed images in the graph below:

**AI-generated images assessed**

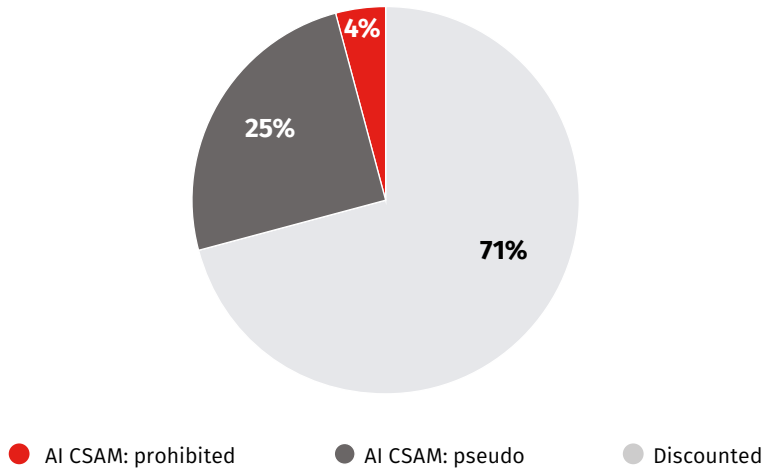



**Figure 4**
Classification breakdown of AI-generated images posted to the forum from March to April 2024.

Source: IWF analysis

**The total proportion of images actioned as criminal was 29% of the unique AI-generated images found on the forum.** 71% of the images on the forum were non-criminal.

Of the criminal images, **over five times as many images were assessed as realistic pseudo-photographs than were assessed as non-realistic prohibited images.** This is close to the results of the previous snapshot, in which the proportion was approximately six-to-one.

Comparing the image assessments between the two studies, relative to the number of AI-generated images posted to the forum over the two periods, including any duplicates in 'discounted' figures, yields the following comparative chart:

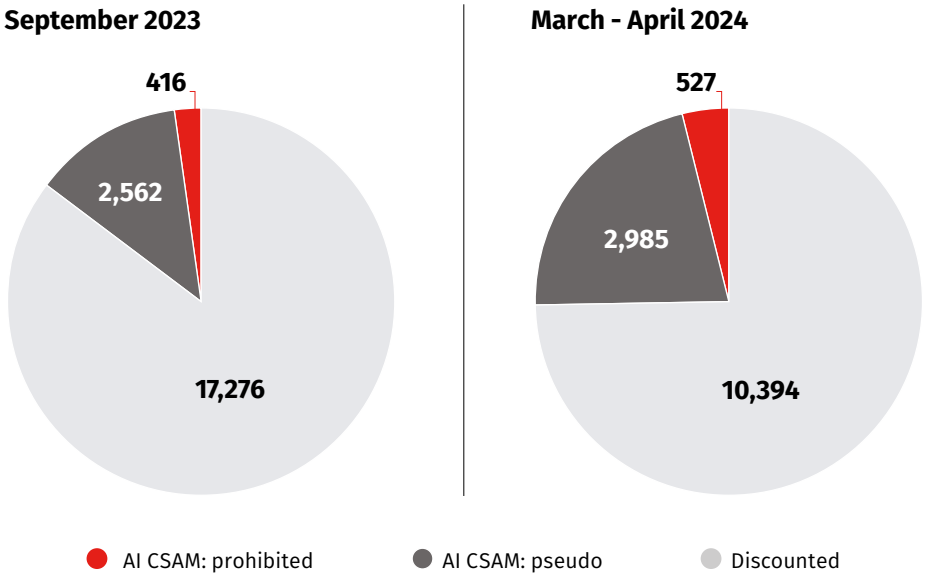



**Figure 5**
AI-generated image assessments compared between the first and second snapshot studies.

Source: IWF analysis

Though the overall number of AI-generated images posted to the forum decreased, among them, a higher number of criminal AI CSAM images were found in this second snapshot.

**The total number of criminal images found on this forum now stands at 6,490.**

Images assessed as indecent pseudo-photographs of children can be sorted by UK Sentencing Council Category and by ages of children. In images in which multiple categories or children are present, the most severe category and youngest age are selected.

The UK Sentencing Council Categories are:

**Category A**
Images depicting penetrative sexual activity; images involving sexual activity with an animal; or sadism.

**Category B**
Images depicting non-penetrative sexual activity.

**Category C**
Other indecent images not falling within categories A or B.

The relative proportions of category and age assessments for AI CSAM (pseudo-photographs) for both snapshots are shown below:

**AI CSAM (pseudo) images by severity**

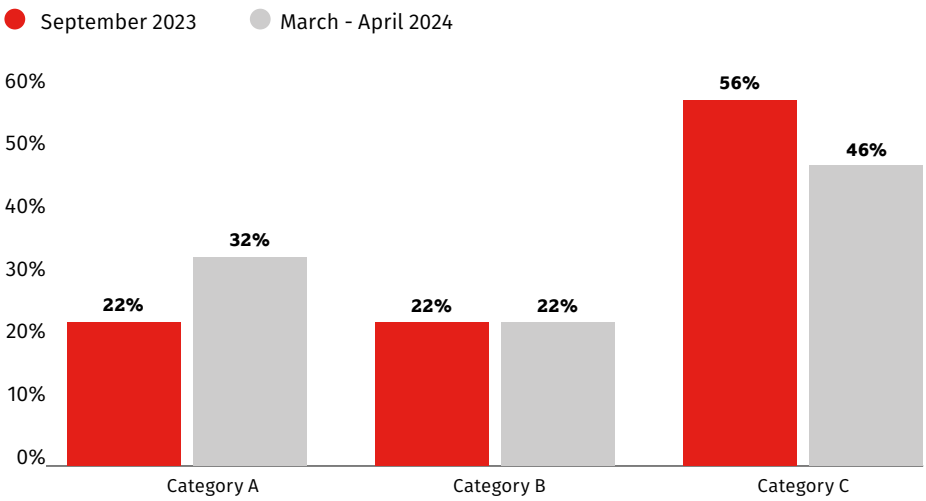



**Figure 6**
AI CSAM (pseudo) images by severity.

Source: IWF analysis

**AI CSAM (pseudo) images by age**

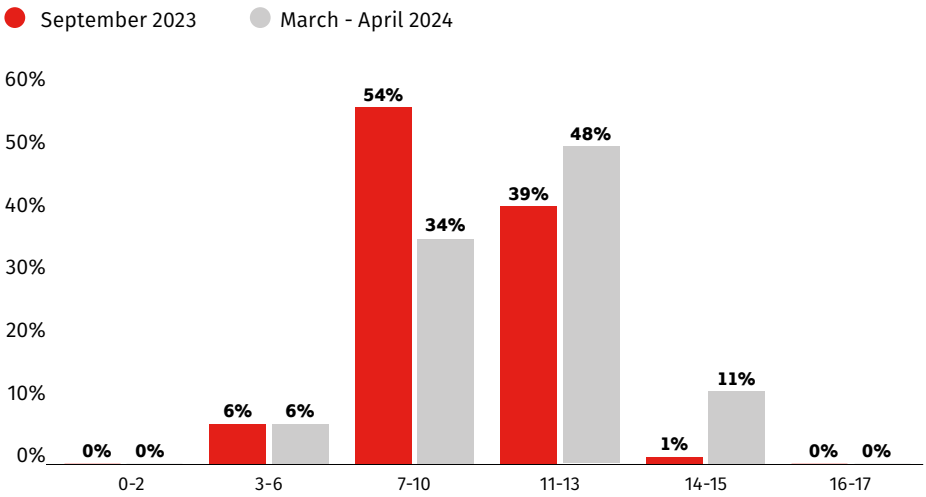



**Figure 7**
AI CSAM (pseudo) images by age.

Source: IWF analysis

Notably, **the proportion of Category A images has increased by 10 percentage points.** This could indicate that technology and expertise has advanced such that perpetrators are experiencing more success generating 'hardcore' scenarios – generally, penetrative sexual activities – complex scenes involving multiple individuals, which image-generating AI has historically had greater difficulties producing accurately compared to scenes involving just one individual.

**Just as in the previous snapshot, female children featured in over 99% of AI CSAM images assessed.**

This second snapshot found 3,512 AI CSAM images out of 12,148 total assessed images. The 8,636 images found to be non-criminal were given reasons for being 'discounted', which can be compared to the discounted reasons for the images assessed in the September snapshot as follows:

| Discounted Reason | September 2023 | % | March-April 2024 | % |
|---|---|---|---|---|
| Age indeterminate | 390 | 5% | **248** | **3%** |
| Child depicted (non-criminal) | 4,340 | 53% | **6,486** | **75%** |
| Adult depicted (non-sexual) | 492 | 6% | **499** | **6%** |
| Non-criminal non-photographic (NPI) | 1,634 | 20% | **651** | **8%** |
| Off remit | 1,274 | 16% | **752** | **9%** |
| **Total** | 8,130 | 100% | **8,636** | **100%** |

In this snapshot, an even lower proportion of non-criminal images assessed – just 8% – were determined to be not realistic enough to assess as a pseudo-photograph (whether that image depicted a child or an adult). These are those images marked 'Non-criminal non-photographic' (NPI) in the table above. This provides some support for claims of increasing realism of AI-generated images over the last six months.

**This snapshot aims to dig deeper into those AI-generated images of children shared in the dark web CSAM forum that were assessed to be non-criminal.**

What was the nature of those images? Could they be mistaken for 'innocent' images of children, or could they be considered sexually exploitative images of children?

IWF analysts were tasked with categorising the non-criminal AI-generated images of children into two categories: exploitative and non-exploitative. These are broadly defined as:

> **Exploitative:** a child depicted such that a reasonable person would consider it sexually exploitative. The image has a (slightly) sexual element but does not meet criminal thresholds.

> **Non-exploitative:** a child depicted in a non-sexual situation. This spans, for example, images of fully clothed children in indoor or outdoor settings, as well as images that may be considered legitimate nudism settings.

It should be noted that the tag 'Child depicted – non-criminal' in the discounted table above – encompassing 75% of the non-criminal images in this snapshot – then indicates that each image in this category did not meet criminal thresholds, but does not necessarily indicate that there was no sexualised element at all.

Even with an exploitative category definition that excludes the many images of children in nudism settings among this set, **a significant proportion (42%) of non-criminal AI-generated images of children were classified as exploitative.** These 2,980 images sometimes featured multiple children – 3,778 children were identified in these images in total. One realistic image depicted 14 children in a sexually exploitative (though non-criminal) context.

Nonetheless, most non-criminal AI-generated images of children, assessed independently – outside the context of their sharing on the dark web – could be considered non-exploitative.

**Discounted AI-generated images of children - non-exploitative or exploitative?**

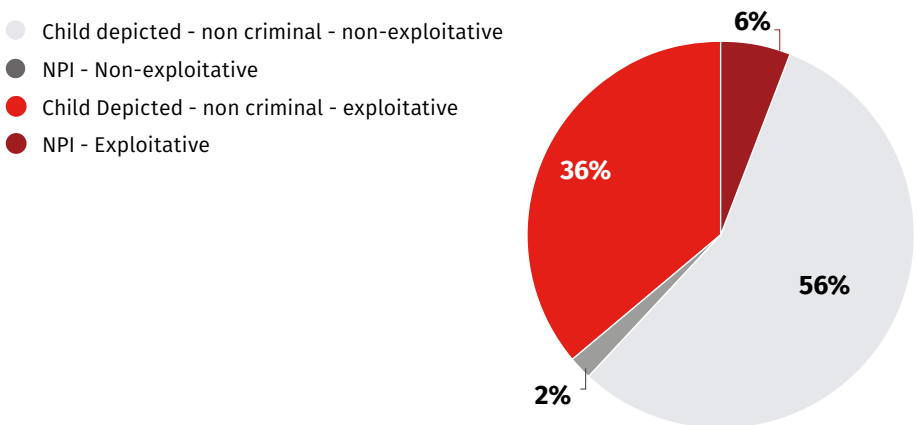



**Figure 8**
Discounted AI-generated images of children - non-exploitative or exploitative?

Source: IWF analysis

For comparative purposes, this snapshot revisited the dark web CSAM forum surveyed for the snapshot in the October 2023 report. This means that the same limitations persist:

- Only one CSAM forum was surveyed.
- The forum surveyed has a general preference towards 'softcore' imagery, and imagery of girls.
- The AI sections of this forum has a few regular 'creators' – though these have changed somewhat from those in the previous snapshot – still, large batches of images assessed originate from the same few perpetrators.

The key findings on images assessed across the two snapshots can be summarised as follows:

- AI-generated images of children, including AI CSAM, already mostly looked like real images of children by the autumn of last year, with 82% of images assessed being realistic enough to be assessed as pseudo-photographs of children (if criminal). **This snapshot found that 90% of images assessed met this threshold, indicating that an even higher proportion of AI-generated images of children look like real images of children now.** It is important to re-emphasise that this proportion will likely never reach 100%, because a small class of perpetrators do not hold realism as a goal – these perpetrators simply prefer non-realistic image styles, like cartoon or anime.
- It remains the case that most AI-generated images of children found are non-criminal, and of those images, most could be classed as non-exploitative. This is further evidence that **there is a large appetite among online CSAM communities for images of children outside scenarios containing explicit sexual activity and outside scenarios that could be considered sexually exploitative.**
- Over time, **both the AI technology itself and its users are getting better at depicting realistic complex scenes involving multiple characters.** This is reflected in the increases in the number and proportion of Category A AI CSAM images.

## Metadata analysis

New for this snapshot, after de-duplication, IWF undertook a scrape of available metadata on all those images shared in this dark web CSAM forum over the 30-day period. Of the 12,148 images assessed, **1,675 images had some 'useful' metadata available.**

The overwhelming majority of those images were shown to have been generated through [redacted], the most popular [redacted] graphical user interface (GUI).

**86% of images shared did not contain any useful metadata.** As discussed in the October 2023 report, this is a consequence of being generated by

open-source models. Processes to strip images of their metadata are usually incorporated by perpetrators into their 'workflows'.

What comprised useful metadata varied between images in this set. This included evidence of the full prompts used; the models used, including foundation models, Checkpoint models and LoRAs; the seed and number of steps in generation; post-processing steps, including upscaling, file conversion, face-swapping (using publicly-available tools), and use of Photoshop. In some cases, combining all this data could be enough to replicate generated images perfectly or almost perfectly.

One important caveat applies to this analysis: it relates to just 14% of assessed images, so, because metadata editing is likely to take place to bulk generations or bulk sets of images, the images included in this analysis are likely to be from the same few sets and the same few perpetrators.

**The most common positive prompt terms related to realistic or photographic images and referred to anatomical features of children.** Other frequent positive prompts described various sexual activities and settings.

**The most common negative prompt terms related to adult sexual or anatomical characteristics, as well as certain racial descriptors.** Other frequent negative prompts described types of body deformities (so-called AI artefacts) as well as those designed to nudge the model towards generating sexually violent or coercive images.

Checkpoint models are large-scale 'base' models that are the bedrock for all image generation. These can be 'foundation' models released officially or they can be models that have been fine-tuned by users. These are often produced and distributed for specific purposes, such as generating pornography.

1,290 images among the set had evidence of use of a Checkpoint model. The remaining 385 images did not contain details of the Checkpoint model used.

**Of these 1,290 images, at least 897 (or 70% of these images) used a Checkpoint model that was publicly available** – freely downloadable from sites like [redacted] or [redacted]. Up to 30% of these images, then, may have used CSAM fine-tuned models. This proportion was roughly the same among the criminal subset of these images that had this data available.

**More LoRA models – smaller fine-tuned models, generally applied on top of Checkpoint models – in evidence among the set, appeared to be CSAM fine-tuned models, including many for generating images of named children.**

SEE RECOMMENDATION #2

For Government

# In summary

**Since the IWF's last AI CSAM report in October 2023, despite increasing levels of legal and regulatory scrutiny, the pace of technological progress has not been slowing.**

With image generation considered close to being 'solved', resources are being poured into trying to solve the next frontier: video generation. We get glimpses of future model capabilities in the 2024 previews of OpenAI's Sora and Google's Veo.

At the same time, focus on specialised text- or image-generating models is to some extent giving way to a focus on models built to be intrinsically multimodal – to work fluently across text, audio, image, video, and code. The future is likely to hold general, all-purpose AI systems that can interpret inputs and produce outputs of all kinds.

AI-generated child sexual abuse and exploitation, just like other types of child sexual abuse, is not limited only to highly technical individuals; to those on the dark web; or to those with extreme or violent sexual urges.

It is a multifaceted threat that encompasses perpetrators of all levels of technical knowledge, including children themselves; sites for distributing, buying and selling AI CSAM on the clear web, as well as on the dark web; people seeking out non-criminal images of children, as well as images of 'hardcore' sexual scenarios.

In this context, there has never been a more urgent need for child safety by design across all the stages of model development and distribution, and among all the players in the AI ecosystem.

While it would be a mistake to assume that closed-source models are watertight without extensive child safety research and testing, open-source models still comprise the main threat in the AI CSAM landscape. If the last few years are a guide, where closed-source models lead, open-source equivalents will inevitably follow. We may yet experience a watershed moment for AI CSAM when fast, malleable generative AI video becomes accessible to the general public.

The AI CSAM videos found in the course of research for this report update, likely created with primitive open-source tools, are the canaries in the coal mine.

With investment in AI safety research, collaborative cross-industry initiatives, and regulation that is adaptable and dynamic, mitigations are possible. And mitigations are needed now – while real harm is being perpetrated against real individuals.

---

For further information on this report update, please email **media@iwf.org.uk**

**DISCLAIMER**

*The images used in this report are screenshots of content available on the clear web and dark web. We've attempted to cite the sources of these screenshots, some of which depict likenesses of famous people or films. These likenesses have been generated by someone submitting prompts to AI models. They are not images of the actors or from the films themselves. This goes some way towards demonstrating the photorealism of images produced by AI models.*

# Glossary

**AI:** *Artificial Intelligence.*

**AI CSAM:** child sexual abuse material that has been generated or edited by Artificial Intelligence.

**Base Model (or Foundation Model):** an AI model, generally those released directly by generative AI companies, designed to produce a wide and general variety of outputs.

**Category A:** a classification of child sexual abuse images depicting penetrative sexual activity; images involving sexual activity with an animal or sadism, as according to the Sentencing Council's Sexual Offences Definitive Guideline.

**Category B:** a classification of child sexual abuse images depicting non-penetrative sexual activity, as according to the Sentencing Council's Sexual Offences Definitive Guideline.

**Category C:** a classification of indecent images of children not falling within categories A or B, as according to the Sentencing Council's Sexual Offences Definitive Guideline.

**Closed-source models:** software whose source code is not released to the public. The public are not able to use, study, change, or distribute the software or its source code to anyone or for any purpose.

**Coroners and Justice Act 2009:** this law criminalises the possession of "a prohibited image of a child". These are non-photographic – generally cartoons, drawings, animations or similar.

**CSAM:** *child sexual abuse material.*

**Dark Web:** the side of the World Wide Web that is not indexed by search engines and requires specific configuration, software, or authorization to access allowing users and website operators to remain anonymous or untraceable.

**Deepfakes:** media (images, videos, or audio) that has been digitally manipulated through AI tools or software to replace one person's likeness convincingly with that of another.

**Diffusion Model:** text-to-image models that add and remove layers of 'noise' to images. Running the 'de-noising' process on random seeds generates 'new' images.

**Generative AI:** a type of machine learning that uses deep learning models to identify the patterns and structures within existing data to generate new content.

**IWF:** *Internet Watch Foundation.*

**LEAs:** *law enforcement agencies.*

**Open-source models:** software whose source code is released under a license in which the copyright holder grants users the rights to use, study, change, and distribute the software and its source code to anyone and for any purpose.

**Open/Clear Web:** the side of the web that is public and viewable by everyone.

**Prompts:** words or short phrases used to describe what you do (positive prompts) or do not (negative prompts) want to see in the image when using generative text-to-image models.

**Pseudo-photograph:** an image (including one generated by a computer) that appears to be a photograph.

**Real CSAM:** child sexual abuse material that has not been generated or edited by AI technology.

**Self-generated content:** when children are groomed, deceived or extorted into producing sexual images and/or videos of themselves and sharing them online.

**Text-to-image model:** a type of machine learning model whose function is to generate images from text prompts.