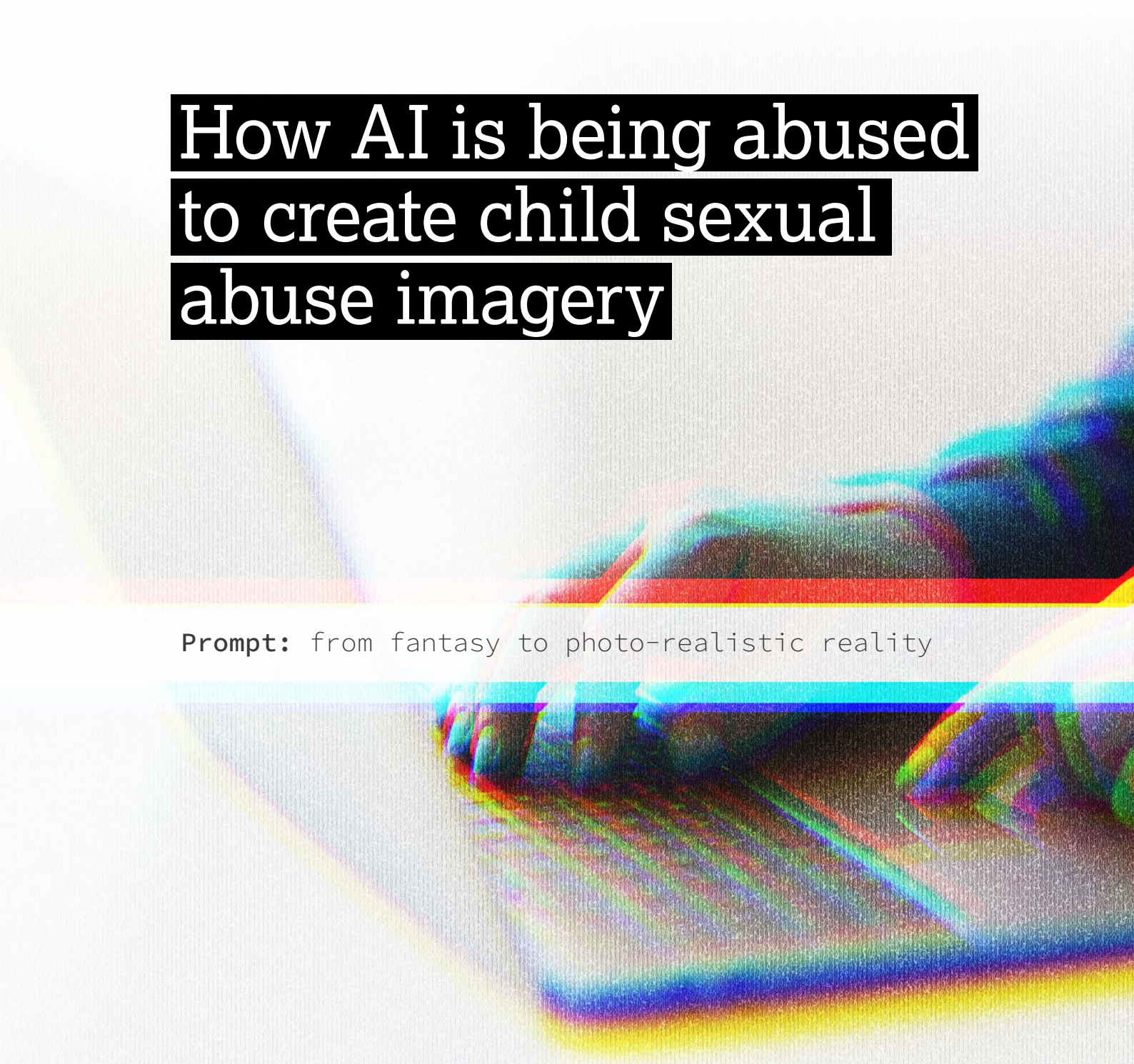
IWF
Internet
Watch
Foundation

October 2023

How AI is being abused to create child sexual abuse imagery



PUBLIC VERSION

Content note

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

This report contains no AI CSAM.

It contains redacted AI adult images.

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 **exactly** 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

Foreword from Susie Hargreaves OBE

Artificial Intelligence brings us to a new frontier in the online world. It promises so much, and we're only just beginning to understand how it can improve our lives, our quality of life, our opportunities. But there is a dark side.

The Internet Watch Foundation (IWF) has always been at the forefront of seeing the abuses of new technology, and AI is no different. What is different where AI is concerned, however, is the speed of development and improvement: When our analysts saw the first renderings of AI-generated child sexual abuse material (AI CSAM) in spring of this year (2023), there were clear 'tells' that this material was artificially generated; backgrounds didn't line up, proportions of body parts were wrong, missing, or clumsy. Half a year on, we're now in a position where the imagery is so life-like, that it's presenting real difficulties for even our highly trained analysts to distinguish. The testimony of perpetrators themselves in dark web forums also tells you want you want to know; there's jubilation that fantasies can be made to order. All you need is the language to tell the software what you want to see.

With the UK Government soon to host an international summit at Bletchley Park on safety within Artificial Intelligence, we thought the time was right to take a more thorough look behind the public reports we have been receiving. What we have discovered confirms our worst fears that this technology is being used to generate indecent images of children.

We are extremely concerned that this will lower the barrier to entry for offenders and has the potential to slow the response of the international community to this abhorrent crime. We have therefore sought to make some recommendations for areas of focus in advance of the important AI summit next month.

We're seeing AI CSAM images using the faces of known, real, victims. We're seeing the 'de-aging' of celebrities and AI CSAM using the likeness of celebrity children. We're seeing how technology is 'nudifying' children whose clothed images have been uploaded online for perfectly legitimate reasons. And we're seeing how all this content is being commercialised.

It's concerning to read some of the perpetrator discussions in forums where there appears to be excitement over the advancement of this technology. What's more concerning for me, is the idea that this type of child sexual abuse content is, in some way, ethical. It is not.

We only need to look to the incredible work of Suojellaan Lapsia (Protect Children) and their Redirection Survey Report where more than half (52%) of the respondents have felt afraid that viewing CSAM might lead to sexual acts against a child; 44% said that viewing CSAM made them think about seeking contact with a child, and more than a third (37%) said they had sought direct contact with a child after viewing CSAM.

It's important that we communicate the realities of AI CSAM to a wide audience because we need to have discussions about the darker side of this amazing technology.

While this report paints a bleak picture, I am optimistic.

We're at the beginning of understanding this technology. Working together, in partnership and collaborating as a sector with industry, law enforcement, Government, and with the right level of funding, we might not be reporting in 12 months' time of how the internet is awash with AI CSAM.

As usual, there is much to do. IWF stands ready to overcome the challenges. What AI creates, I'm hopeful AI can solve.

Susie Hargreaves OBE

CEO

Recommendations

FOR GOVERNMENT:

1 To explore at the Government's forthcoming AI Summit the challenges for dealing with AI CSAM including the need for alignment internationally on how this content is treated in different jurisdictions and secure commitment to ongoing collaboration from international governments and stakeholders.

2 For the Ministry of Justice to commission a review of the laws that apply to the removal of this content online to ensure they are fit for purpose to tackle the threat of AI CSAM. This includes ensuring the exchange of "hints and tips" and "paedophile manuals" on how to generate this content are made illegal.

3 To consider an extension of the IWF's remit to be able to scrutinise the datasets on which these technologies are trained.

FOR LAW ENFORCEMENT AND REGULATORS:

4 To Ensure the College of Policing training course is updated to cover AI CSAM, and clear guidance is issued to police graders on how to process this imagery.

5 To ensure there is proper regulatory oversight of AI models before they go to market or are made open-source and ensure appropriate risk mitigation strategies are in place. For closed source models, protections must be in-built.

FOR TECHNOLOGY COMPANIES:

6 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.

7 That search services should de-index links to fine-tuned AI models known to be linked to the creation of AI CSAM.

8 To carefully consider the content moderation challenges AI CSAM creates in terms of prioritisation and the mixed nature of AI CSAM with real CSAM.

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

Executive summary

Child sexual abuse images generated using artificial intelligence is a new and growing area of concern.

The key findings of this report are as follows:

In total, **20,254 AI-generated images were found** to have been posted to one dark web CSAM forum **in a one-month period.**

Of these, **11,108 images were selected for assessment by IWF analysts.** These were the images that were judged most likely to be criminal.

(The remaining 9,146 AI-generated images either did not contain children or contained children but were clearly non-criminal in nature.)

12 IWF analysts dedicated a combined total of 87.5 hours to assessing these 11,108 AI-generated images.

Any images assessed as criminal were criminal under one of two UK laws, as described in section 5. These are:

- 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.

2,562 images were assessed as criminal pseudo-photographs, and 416 assessed as criminal prohibited images.

Other findings:

1. AI-generated content currently comprises a small proportion of normal IWF activities, though one of its defining features is its potential for rapid growth.
2. Perpetrators can legally download everything they need to generate these images, then can produce as many images as they want – offline, with no opportunity for detection. Various tools exist for improving and editing generated images until they look exactly like the perpetrator wants.
3. Most AI CSAM found is now realistic enough to be treated as 'real' CSAM. The most convincing AI CSAM is visually indistinguishable from real CSAM, even for trained IWF analysts. Text-to-image technology will only get better and pose more challenges for the IWF and law enforcement agencies.
4. There is now reasonable evidence that AI CSAM has increased the potential for the re-victimisation of known child sexual abuse victims, as well as for the victimisation of famous children and children known to perpetrators. The IWF has found many examples of AI-generated images featuring known victims and famous children.
5. AI CSAM offers another route for perpetrators to profit from child sexual abuse. The first examples of this new commerciality have been identified by the IWF.
6. Creating and distributing guides to the generation of AI CSAM is not currently an offence, but could be made one. The legal status of AI CSAM models (files used for generating images) is a more complicated question.

Introduction to this report

This year, the Internet Watch Foundation (IWF) has been investigating its first reports of child sexual abuse material (CSAM) generated by artificial intelligence (AI).

Initial investigations uncovered a world of text-to-image technology.

In short, you type in what you want to see; the software generates the image.

The technology is fast and accurate – images usually fit the text description very well. Many images can be generated at once – you are only really limited by the speed of your computer. You can then pick out your favourites; edit them; direct the technology to output exactly what you want.

These images can be so convincing that they are indistinguishable from real images.

The most convincing AI CSAM images, then, can be called photorealistic. For IWF analysts, looking at this sort of AI CSAM is exactly like looking at 'real' images of the sexual abuse of children. Except these images have been generated by algorithms.

Images show the rape of babies and toddlers; famous pre-teen children being sexually abused; BDSM (bondage and discipline, dominance and submission, and sadomasochism) content featuring tweens and teenagers. And more.

Effectively articulating the criminality of AI CSAM can be a challenge – there are groups who seek to lessen the severity of these images: they 'don't have real children', or 'don't hurt anyone'.

UK law, however, is clear: AI CSAM is criminal.

Images that are not realistic – that appear like cartoons or drawings – are "actionable" by our analysts (criminal, and therefore able to be removed from the internet under UK law) under laws on prohibited (non-photographic) images of children.

Images that are realistic – that appear to be photographs – are actionable under laws on indecent pseudo-photographs of children. (For precise laws, see section 5).

Amid all the focus on realism, photorealism, and hyperrealism, and complex debates about legality – simply stated – this technology allows perpetrators to generate dozens, even hundreds of child sexual abuse images at the click of a button.

Crucially, you can download AI technology (at just a couple of gigabytes) and run it on your device offline. So, once you have the technology, you can generate as many child sexual abuse images as you like – 'in the dark', with little or no risk of detection.

The genie is out of the bottle. Offline child sexual abuse image generation is our reality.

What does this mean for IWF?

Currently, AI CSAM represents a small portion of the vast numbers of 'real' CSAM we find. (Over 255,000 webpages last year, representing hundreds of thousands or even millions of images.) Time will tell whether this trickle becomes a flood.

Some websites have been set up that are dedicated to sharing AI-generated images, but we are also starting to see AI-generated images mixed in with 'real' images. These images can be especially difficult for analysts to detect as AI-generated – to tell 'real' from 'fake'.

As the technology continues to improve, and perpetrators generally get better at generating realistic images, this challenge will only get harder.

These websites are still reported, and removal is pursued. UK law is clear, but if websites lie in other jurisdictions, removal can be more complicated.

IWF tags all these images to identify them as AI-generated, which helps law enforcement and victim identification (VID) efforts.

Questions remain. How can safeguards be built into this technology, even if offline image generation is possible? Is AI image detection possible and practicable? Is the law fit for purpose, or should it be changed? Will mass quantities of AI-generated images enfeeble hash lists?

Lots of discussion about the risks of AI – discussion that spurs moves to regulate AI companies – centres around hypothetical or long-term risks like creation of synthetic viruses, cyberattacks or, at the extreme, the risks in creating a 'superintelligence', or postulated artificial general intelligence (AGI).

AI CSAM is different because it is happening now. Images are being shared online now. It is a current problem that requires action.

At the same time, solutions developed and implemented now have the potential to mitigate this problem.

With all technological advance comes benefits as well as risks. Though this report focuses on current abuse of AI technology to generate CSAM, it is important to bear in mind the widespread potential for benefits from AI across society, from applications in science, research, and healthcare, to applications in the creative and entertainment industries.

Nonetheless, left unchecked, this technology will cause harm to children.

It harms known victims of child sexual abuse, whose likenesses are being used to generate more images of them in new scenarios.

It harms new victims of child sexual abuse, whose potential investigators might spend time and resources pursuing the rescue of children who turn out to be virtual characters.

These images provide new possibilities for perpetrators to use to groom and coerce children. They even allow the most technically proficient perpetrators to make money from abuse.

And this is the worst in terms of quality of output that AI technology will ever be. It only has the potential to get better: to produce more lifelike images; to better enable the grooming and abuse of children.

Overall, AI CSAM poses a significant risk to IWF's mission to remove child sexual abuse material from the internet.

What is the IWF and why has it produced this report?

The IWF is a not-for-profit organisation, funded by tech companies, government, global funders and the public, whose remit is to remove CSAM from the internet.

The IWF Hotline, which finds, assesses, and seeks removal of this criminal content, has two main sources for its work: reports from the public (and external partners), and proactive searching for content.

This year, the Hotline has received its first reports of AI CSAM, mostly from members of the public. Reporting numbers were – and remain – small relative to the number of other CSAM reports.

Nonetheless, subsequent proactive searches for AI CSAM found widespread evidence for a large and growing problem. Images and intelligence obtained from these proactive searches have informed IWF media pieces that have raised awareness of this problem and the enormous potential for abuse. Consultations with government and civil society about how to address this problem are ongoing, and discussions with industry in the early stages.

Remit and scope of this report

The IWF has the remit to investigate publicly accessible areas of the internet, but not peer-to-peer networks (including end-to-end encrypted chats) or most content that is hidden behind payment barriers. These areas fall under the remit of law enforcement.

This report is informed by intelligence shared by law enforcement partners relating to AI CSAM in these publicly inaccessible areas, but concrete discussions and examples in this report relate to content that has been found on the clear web and dark web. This report should be read in conjunction with reports from law enforcement partners that discuss AI CSAM in inaccessible areas.

AI CSAM is related to other important AI topics and themes that are out of scope of this report. These include, among others, intellectual property and copyright questions; generation of sexual images of non-consenting adults; misinformation and disinformation; bias, including questions of AI sexism and racism; and using AI to generate terrorist, violent or other illegal material.

Somewhat out of scope of this report are uses of generative AI to coerce and groom children beyond the generation of AI CSAM. This includes, for example, use of Large Language Models (LLMs) in offending against children, or use of AI chatbots and their role in offending pathways (as highlighted by law enforcement partners).

Notes on terminology

Terms in the AI field are used variously, and often overlap. This report clearly defines the terms it uses throughout. The most important are defined in section 4.

This report uses the term 'AI CSAM' to refer to criminal images or videos of children that have been generated or edited by AI technology. This software is most likely to be text-to-image in nature but could also take other forms (more details in sections 4-5).

To clearly distinguish CSAM content that is not generated or edited by AI technology, this report uses 'real CSAM'. This term should not be taken to diminish the severity or criminality of AI CSAM.

This report uses the term 'perpetrator' over 'offender' or 'criminal' to reflect the IWF's role as a non-law enforcement agency – to avoid overstepping IWF remit by assigning criminality to individuals.

The word 'generate' is preferred over terms like 'create', 'make', or 'produce' to avoid problems with assigning creative agency to text-to-image software, and to emphasise that this software is a neutral tool.

Outline and guide for readers

This report will begin with an outline of generative AI and how it is used to generate child sexual abuse images. The following sections will describe the technology and tools being used.

Given the intention of this version of the report to be placed into the public domain, information will be deliberately limited so as not to resemble a guide on how to create this material.

This report will then focus on actual cases of online AI CSAM – where it is found; how much there is; relevant questions and issues.

Finally, this report will briefly consider detection and enforcement, including impacts on law enforcement agencies (LEAs) and analyse whether legislative gaps exist in this area.

4

Generative artificial intelligence

Though artificial intelligence is a decades-old research field in computer science, it experienced a turning point in November 2022 with a dramatic increase in public and media attention following the release of the text-generating program ChatGPT.

Simply put, the aim is to make intelligent computer programs. To achieve this, AI systems are trained on data within which they can draw connections and look for patterns. An iterative learning process takes place, guided by a combination of human feedback (supervised machine learning) and/or algorithmic feedback (unsupervised machine learning).

Depending on the input dataset and the type of machine learning used in this process, AI systems can have different purposes. The AI systems described in this section are trained by a process called deep learning, which is a type of machine learning that is loosely modelled on the human brain – using artificial neural networks. These deep learning systems are trained on huge datasets scraped from the internet.

Systems that are trained on huge quantities of text and whose function is to generate text – used, for example, in online chatbots – are large language models (LLMs). These models are renowned for their natural language processing abilities, 'understanding' and interpreting human language. Leading LLMs include:

- ChatGPT, developed by OpenAI
- PaLM, developed by Google
- LLaMA, developed by Meta
- Claude, developed by Anthropic

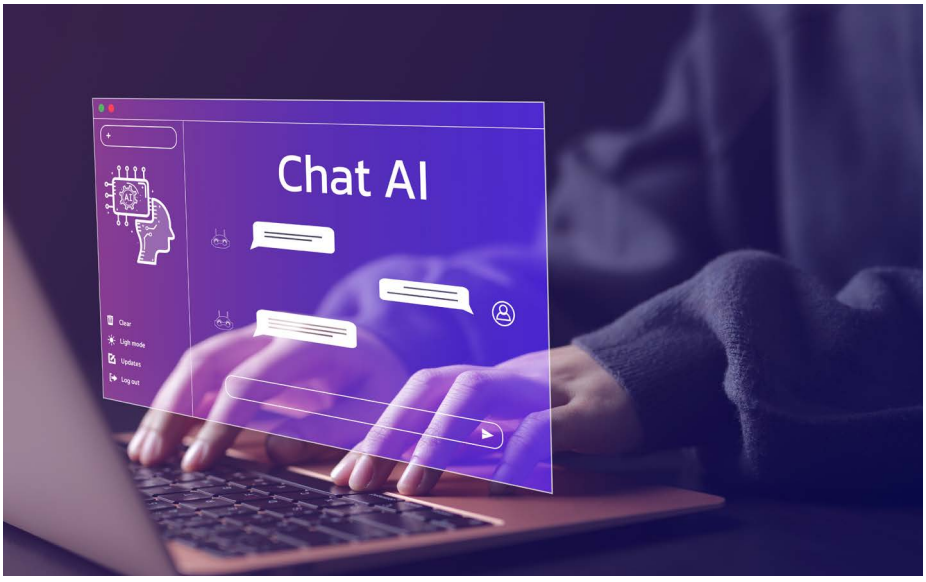


Figure 1
Large language models (LLMs) generates naturalistic text

AI systems may extend into more modalities than just text: speech and audio, video, code, 3D modelling data, or others. A system is *multimodal* if it combines these modalities.

Systems that are trained on huge quantities of tagged images (in other words, images with key descriptive terms attached) and whose function is to generate images are AI image generators or text-to-image models. These systems are, therefore, multimodal – they 'translate' from one modality to the other. Leading text-to-image models include:

- Midjourney, developed by Midjourney, Inc.
- DALL-E, developed by OpenAI
- Stable Diffusion, developed by Stability AI

The versions released by these companies are called base models or just models. These can produce highly realistic images.

Stable Diffusion 2.0 **Midjourney V4**

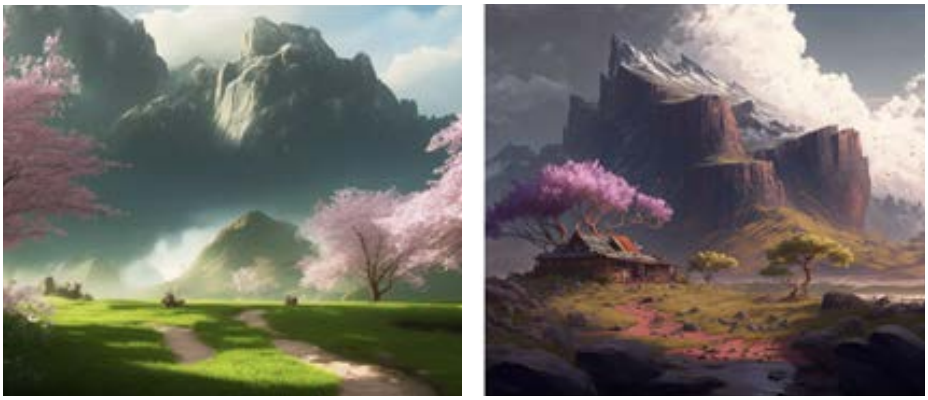


Figure 2
A comparison of two cinematic images produced using the same prompt on two text-to-image models

Source: Medium (Jim Clyde Monge)

Image generation

Image generation also saw a rapid rate of progress last year with releases like Midjourney V4 in November. This version represented a step change in the quality of AI image generator available.

New releases of these models incrementally improve the quality of images that they can generate. Midjourney now has a v5.2, released in July; Stable Diffusion has a new version called SDXL, also released in July; DALL-E 3, which some consider the most impressive of all these models, was released in late September.

Figure 3
A comparison of versions of Midjourney using images generated using the same prompt

Source: Reddit (r/midjourney) / Youtube (Curtis Pyke)

Various types of text-to-image models exist, including generative adversarial networks (GANs) and variational autoencoders (VAEs), but all current cutting-edge AI image generators – including Midjourney, Stable Diffusion and DALL-E, are diffusion models.

To train a large-scale diffusion model requires a vast dataset of images that are scraped from the internet and then labelled with descriptive words or phrases – the type of text that will later be used for prompting new generations.

Generally, the collection and necessary tagging of these images is outsourced to other organisations:

- Midjourney and DALL-E use a diffusion model conditioned on contrastive language-image pre-training (CLIP) image embeddings. CLIP is trained on 400 million pairs of images, scraped from the internet, with text captions.
- Stable Diffusion uses a dataset called LAION, developed by a group of European researchers, which has 2.6 billion English language-tagged images within a 6 billion image dataset. Various Stable Diffusion versions were trained on samples of millions of images taken from this vast dataset.

Diffusion models work by adding random Gaussian noise (or simply 'noise') to images, then running the reverse process, 'denoising' step-by-step to reconstruct each image. 'New' images are generated by changing the noise before running the reverse, learned denoising process again.

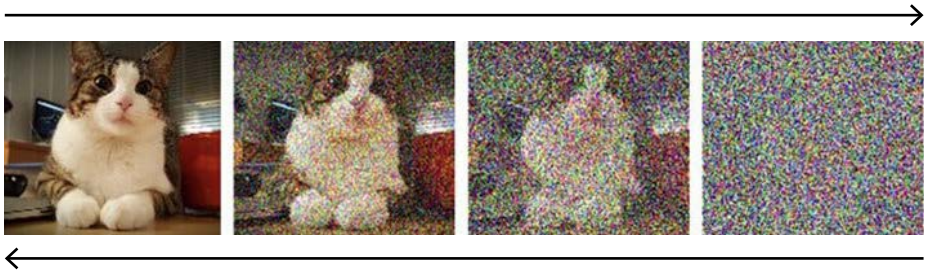


Figure 4
Simplified diagram shows the process of adding and removing Gaussian noise to images (indicated by the arrows) in the diffusion process

Source: Nvidia

During the reverse, denoising process, layers of noise are removed to generate an image. A diffusion model can be stopped at any point along this reverse process, but would output a noisy, 'fuzzy' image if stopped too early.

A key feature of diffusion models is that they are interpolative. They draw connections within the training dataset and can generate new images within the semantic bounds of this dataset.

Diffusion models are large-scale and can generate detailed, high-quality images.

Because these models can generate photorealistic images, much discussion centres around AI image detection: how to tell when an image has been generated by AI. This discussion is fuelled by recent media cases in which people have been 'fooled' by AI-generated images and gives rise to concerns about media authenticity and misinformation.

Some momentum exists for establishing a set of standards in relation to digital watermarking of AI-generated images. The idea is to embed the fact that an image is AI-generated into image metadata, to increase trust in media and reduce potential for misinformation. AI companies may also want to tag images generated by their systems so that AI-generated images are excluded from future training datasets.

No common standard for digital watermarking currently exists; the main two being adopted by industry are IPTC and C2PA. A new Google technology called SynthID adds a watermark to individual pixels in images.

Figure 5
Google's SynthID image watermarking is resistant to some image editing, shown in the variations on the original butterfly image

Source: Google DeepMind

With no common standard, it is unsurprising that no fully reliable AI image detection tool exists, even if some tools claim high accuracy.

Questions of digital watermarking and detection are further complicated by the fact that image metadata can in principle be edited in and edited out of images. AI-generated images, therefore, could have AI-identifying metadata removed, or real images have AI-identifying metadata added.

Implications of difficulties of telling 'real' CSAM from AI CSAM are discussed further in sections 9-10.

Closed-source and open-source models

If a technology company keeps the code that comprises its software secret, not releasing it to the public, the software is closed-source. If it decides to release the code, it is said to be making it open-source.

An AI company may seek to keep a model closed source for commercial reasons – to avoid sharing development secrets with rivals. It may want full access to data concerning its users and their interactions with the model or favour the increased content moderation options for closed-source models (see section 5 for further details).

In contrast, an AI company may release the code for a model because it believes in open access and the democratisation of technology. It may be attracted to the opportunity for a community of developers, able to share all relevant information and code, to collaborate and add improvements and edits to base models.

Though this report will focus on misuse of open-source models, this is a complex debate that has no simple solution as condemning those companies that make their models open-source. There are risks and benefits to both approaches.

Midjourney and DALL-E are both closed-source models (both are cloud-based models). Stable Diffusion is an open-source model.

Overview: from AI images to AI CSAM

This year has seen a leap in the level of detail and realism in AI-generated images. If AI models can now generate photorealistic images, they can generate photorealistic images of children. If AI models can generate pornographic images, they can generate photorealistic CSAM.

Content moderation

Text-to-image AI companies take different approaches to the question of permitted content – what they allow their models to generate. In general, they seek to disallow restricted content, like violence or pornography. This could be as part of self-regulatory efforts, or as a pre-emptive move to avoid greater moves to regulate the sector.

SEE RECOMMENDATION #6

For Technology Companies

Terms of use provided for the main models are:

DALL-E (OPENAI)
Our content policy does not allow users to generate violent, adult, or political content, among other categories. We won't generate images if our filters identify text prompts and image uploads that may violate our policies

MIDJOURNEY (MIDJOURNEY, INC.)
Do not create images or use text prompts that are inherently disrespectful, aggressive, or otherwise abusive. Violence or harassment of any kind will not be tolerated. No adult content or gore. Please avoid making visually shocking or disturbing content. We will block some text inputs automatically.

STABLE DIFFUSION (STABILITY AI)
You agree not to use the Model or Derivatives of the Model: - In any way that violates any applicable national, federal, state, local or international law or regulation; - For the purpose of exploiting, harming or attempting to exploit or harm minors in any way; - To generate or disseminate verifiably false information and/or content with the purpose of harming others... [continues]

Broadly, content moderation methods for text-to-image models can be divided into two categories:

1. **Restricting training data.** These models are interpolative – they can generate only those things to which they have been exposed. If a model is not exposed to pornography, for example, it will not be able to generate pornography (except by combining various concepts about which it does know – a necessarily limited approach).

If AI models can generate pornographic images, they can generate photorealistic CSAM.

2. **Banning prompts.** By restricting the terms that can be used to generate images, perhaps using a keywords list, concepts that contravene content policies can be excluded from possible generations.

Closed-source models, whereby the company has full control over model training and use, can employ both methods for content moderation. DALL-E and Midjourney use both methods to a high level of effectiveness in the CSAM domain, for example.

Open-source models can attempt to employ these methods, but encounter problems with each, since the code is necessarily editable, and base models can be fine-tuned – trained on further images.

How does Stable Diffusion, the leading open-source text-to-image model, seek to enforce its terms of use?

Since Stable Diffusion v2.0, released in Autumn 2022, pornographic (NSFW – not safe for work) content has been excluded from training datasets. Asked about why Stability AI was taking this approach, Emad Mostaque, CEO, reportedly referenced images that

> ***"could cause legal troubles for all involved and destroy all this. I do not want to say what it is and will not confirm for reasons but you should be able to guess."***

One AI CSAM perpetrator on a dark web forum explains:

> *"Stable Diffusion 2.0+ used a different, much more filtered data set so it's much harder to make NSFW content, not just CP ['child pornography'] but any kind of nudes/porn."*

Nonetheless, Stability AI cannot in practice prevent its models from generating images that would contravene its terms of use above.

AI pornography

Generating pornography, then, is difficult or impossible through some models, and possible with early versions of others. This is because some closed source models cannot generate pornography either because they lack the necessary training data or because users are disallowed from prompting the generation of pornography. Base versions of open source models which contain pornography in their training data, and have no prompt restrictions, allow for pornography generation.

Another route to generating AI pornography is through websites that are dedicated to providing this service – these often use built-in models. These sites seem to have been increasing in number this year; the IWF Hotline increasingly receives public reports relating to content found or generated on them.

Another route to AI-generated pornography is found through services designed for 'nudifying' images. A user uploads an image of a clothed individual; the model outputs an interpretation of the individual without clothes. Sites also exist dedicated to providing this service.

Such examples demonstrate the significant overlap between discussion about AI-generated images and other kinds of 'fakes': 'deepfakes', edited content that may involve use of generative AI, and 'shallowfakes', which include content edited using editing software.

The prevalence of pornography on the internet reflects high demand among consumers. This means firstly that there is lots of content, easily accessible, for use in AI training datasets; secondly, that there is high demand for bespoke or custom-made pornography featuring preferred individuals, styles, positions, and activities. In this context, the fact that such a large proportion of online text-to-image content is pornographic is unsurprising – as is the growth of AI pornography communities on some social media sites.

In summary: through any or a combination of these approaches, photorealistic AI-generated pornography can be obtained. In principle, there is no technical barrier to generating images of younger individuals, including children.

Viewing and assessing AI CSAM

AI CSAM is criminal – actionable under the same laws as real CSAM. These are:

- **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.

IWF analysts assess each AI-generated image to determine whether it meets the threshold for criminality under one of these Acts. The key criterion for classification as criminal under the PoC Act is that the image "appears to be a photograph".

Proving whether an image is AI-generated is not an evidential requirement for prosecution under the PoC Act – it only needs to look like a photograph and be an indecent image of a child.

Under which law - the PoC Act or the CJA -does AI CSAM tend to fall? Answering this question can be a real challenge for IWF analysts:

IWF analysts are able to action AI CSAM that is criminal under these two laws. However, the increasing realism of AI CSAM has presented significant moderation challenges for our analysts and there are different attitudes internationally to non-photographic and computer-generated (CG) imagery. This means that removal of AI CSAM from the internet may be slower and more complex than removal of real CSAM from the internet.

SEE RECOMMENDATION #2

For Government

Why is AI CSAM increasing in realism? This change can be ascribed to several factors: improved AI models; growing communities sharing AI content, tools and tips; improved technical ability in general among AI CSAM-sharing communities.

As part of this report, IWF analysts assessed thousands of AI-generated images. Their thoughts and comments, including on making these assessments, are collected in sections 9 and 11.

Overview: technology and tools

Choosing a model for pornography

Individuals intending to generate AI pornography for the first time are likely to try a low-effort, easy-to-use website.

The gap between easy-to-use models and difficult-to-setup and time-consuming models has led to the growth of websites dedicated to providing a simple pornography generation service.

These sites usually offer a selection of options for features of the image: age; body features; position or activity; setting; and more. Others allow for positive and negative prompts and multi-image generation.

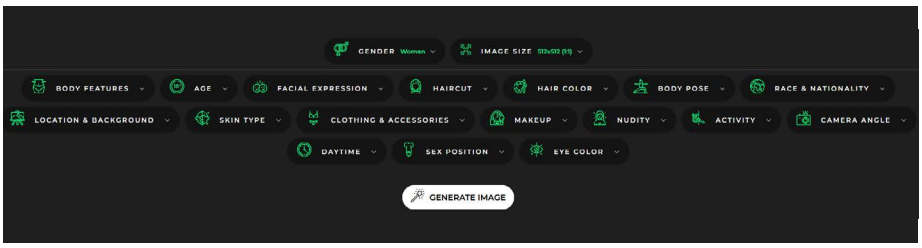


Figure 6
Options for an online AI pornography generation tool allow customisation of various elements of the generated image

Source: Pornderful
Author's screenshot

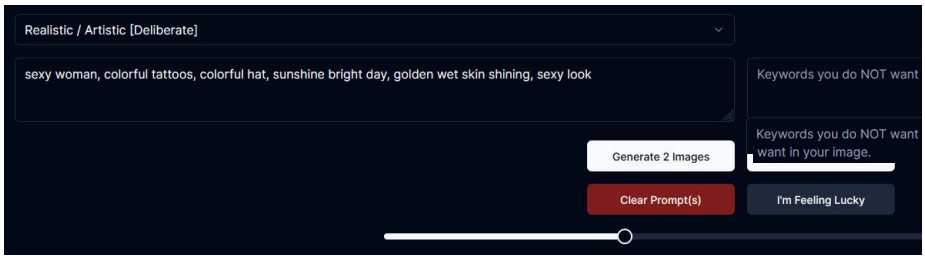


Figure 7
A different AI pornography generation tool allows positive and negative prompts, like Stable Diffusion

Source: sexy[.]ai
Author's screenshot

As outlined in section 5, other websites offer 'nudifying' services. These can be as simple as just requiring the user to upload a photo, then clicking a button.

SEE RECOMMENDATION #6

For Technology Companies

Terms of use for all these websites usually prohibit underage content, but how these terms of use are enforced is unclear. Anecdotal evidence that suggests that some of these measures are working is provided in section 7.

Refrain from sharing any content that is offensive, induces violence or crime, or is seen as threatening, defamatory, or harassing. Posting illegal content, including but not limited to child pornography or unauthorized ('revenge') pornography, is forbidden.

- **Do not post realistic (minors) • Do not post questionable images of minors (both anime and realistic)**

Another website claims to take measures to prevent generation of AI CSAM, though exact measures are not named:

As long as your country allows porn, then you are good to go. But it's worth to mention that ai child porn is forbidden on [website] even if it's allowed in your country. We spend lots of our time to prevent child abuse ai content generated from our site. You will get banned if you try to generate it.

Media attention on such websites and services is increasing with a growing number of stories about production of nonconsensual deepfake images of adults and children.

Despite some barriers to entry to generating adult images with an open source model, a growing community exists. Individuals may be attracted by the infinite range of possibilities on offer; by a technically adept community that shares images, tips, and models; or by the option to generate more graphic and extreme content, including criminal categories of content that would be disallowed by online closed-source models (including violence, animal abuse, gore, rape, or child sexual abuse).

There are sites designed for sharing fine-tuned models for open source AI software. The prevalence of not-safe-for-work (NSFW) content throughout one site reflects the high demand for use of a particular open source model to generate pornography. Popular models available on one site include those for generating lifelike images of various celebrities; models for turning images into distinct anime styles; even models for making individuals in generated images look younger. A combination of an NSFW prompt or model with any of these model types (and many more) is not only possible but commonplace.

7

AI CSAM: technology and tools

This section details the technology used to generate AI CSAM. For discussion about AI CSAM images, including conversations between perpetrators about realism and whether these images are 'satisfying', common features of these images, and commerciality, see section 9.

Verbatim comments from perpetrators originate from various sources. Many are taken from investigations on forums on the dark web – discussion sites where real CSAM is shared; help and advice offered; stories and anecdotes exchanged. Forums have sections or subsections and threads on which posts (text, images, links, or a combination of these) are made. For more information about the spread of AI CSAM on these forums, see section 8.

Choosing a model for CSAM

Just as it is likely that individuals coming to AI-generated pornography for the first time will try low-effort routes, it is likely that perpetrators looking to generate AI CSAM will try similar routes. Evidence that those AI pornography websites that ban the generation of underage characters are frustrating low-tech perpetrators can be found in this comment from a dark web forum user:

> *"Looks great but how are you doing this without getting flagged? I tried to do this on a random AI generator and I couldn't use certain words like "nude" and "teenager"... I tried things in combination and it would flag me as inappropriate words"*

Browser based models, says another user,

> *"have word filters and negatives that prevent most prompts anyway."*

Perpetrators frustrated by these routes will likely move on quickly.

It is accepted among AI CSAM communities that trying to use online services to generate CSAM entails huge security risks.

Other users discuss the potential of the cloud-based, closed-source models for the same reason, these routes are inaccessible:

> *"If only there were [name] without censorship, and it would be possible to make models there"*

> *"if they ever release a console version of [name] that is disconnected from the Server, the stuff you could make is beyond crazy"*

If all these routes are inaccessible, how do perpetrators start?

The overwhelming consensus names one open source solution as the method for generating AI CSAM:

> *"Most are using [name], an AI art generation tool. In order to get it to create on topic images you need to run it locally rather than using online tools"*

> *"[name] created a user interface (WebUI) to make it easier to set up and use [name]. This is what most people are using.His project is relatively simple to set up locally and there are plenty of guides and YouTube videos available."*

A guide to generating AI CSAM with one particular model from early 2023 was widely shared in dark web forums.

Much of the most realistic AI CSAM found in investigations for this report used fine-tuned CSAM models.

There are CSAM models that are well-known among AI CSAM communities – reputed for enabling realistic generation of certain CSAM scenarios, children, or child characteristics. These models are updated – new releases made – by technical experts in the community.

Models fine-tuned on CSAM are not illegal or criminal under UK law. Further discussion on legality of fine-tuned CSAM models is found in section 11.

SEE RECOMMENDATION #2

For Government

Generating images of known victims and famous children

CSAM fine-tuning often uses datasets that feature a particular child individual – usually a known victim of child sexual abuse, or a famous child. This is because, for both these categories, large enough image sets exist to train AI models. The former category has the additional advantage of containing pornographic images, so the output model may not need to be combined with other fine-tuned models for pornography.

The IWF has been aware for a long time of the tendency among perpetrator communities to 'collect' content featuring their preferred child sexual abuse victims. Perpetrators have 'favourite' victims; share content featuring that victim; and look for more.

Now, perpetrators can train a model to generate as many new images of that victim as they like.

These models are comparable to 3D models insofar as they aim to reproduce the likeness of that victim as closely as possible but retain the flexibility to transpose generated character(s) into any setting; any scenario; any type of activity.

The same holds for celebrity children – just as the IWF has for a long time seen many examples of 'shallowfake' and deepfake images featuring these well-known individuals, now the IWF is seeing entirely AI-generated images produced using fine-tuned models for these individuals.

An increasing number of AI CSAM shared on dark web forums features known victims and famous children. Many of these are requested by other users – of the type 'Can you make a model of X' or 'can you make images featuring X' – produced to specification. This includes transposing victims of illegitimate child modelling operations (Category C images) into new, Category A scenarios.

> *"Someone asked for 5-8 year olds in lingerie over in the Request section. So..."*

On another forum, a guide to creating models using personal CSAM datasets has been shared.

What if perpetrators lack the necessary datasets? Various threads found on dark web forums shared large sets of faces of known victims for creating deepfakes or for training AI models. Indeed, one thread was called, 'Photo Resources for AI and Deepfaking Specific Girls'. Perpetrators discussed how to gather images and choose which to use for fine-tuning.

In another vein, evidence has been found of perpetrators creating virtual 'characters' – entirely AI-generated children whose models may have been trained on real children but do not resemble real children – comparable to 'virtual celebrities' or 'VTubers' – and sharing packs of their images.

How many fine-tuned CSAM models are being shared? Obtaining a definite number is impossible – and the IWF does not have the remit to test these models, even if they were all downloaded – but just one forum had just over 100 posts claiming to share these models. Threads with the most popular CSAM models on the forum had tens of thousands of views. For more information and statistics on AI CSAM prevalence, see section 8.

Generating AI CSAM: summary

Photorealistic AI CSAM can be generated (on-device and at scale). There is also technology and techniques to be able to further refine the imagery.

Fine-tuned models allow for the (bulk) generation of images featuring known victims and famous children. There is widespread evidence for perpetrators sharing these models and requesting new ones for their favourite victim(s).

The best AI CSAM looks like real CSAM. In the words of one impressed viewer:

> *"I doubt anyone would suspect these aren't actual photographs of an actual girl."*

8

AI CSAM: prevalence

IWF reports

The IWF has been receiving a small number of reports of AI CSAM from members of the public. (See Forum Snapshot Study for details of an IWF proactive investigation into AI CSAM.) Most are **not "actionable"**. (An "actionable" image is one which breaches UK law and therefore, our analysts will work with relevant partners to have it removed from the internet.)

Communities sharing AI adult pornography have been growing in these places.

Statistics for all reports containing generative AI content (criminal and non-criminal), accurate at time of writing, are provided below.

AI identified in report	93-104 reports
AI identified in actionable report	24 reports
AI CSAM identified in actionable report	15 reports

The first statistic reflects some uncertainty owing to the removal of webpages before manual review.

The key statistic showing the encroachment of AI CSAM into typical IWF activities is the third row above: **15 websites** that were actioned as potentially criminal under UK law that contained, in whole or in part, actionable AI CSAM.

It is unsurprising that a low proportion of reported sites are actionable – this reflects IWF external reports in general (just 12% of all external reports were actionable in 2022, for example. This figure rises to 26% when you include duplicate reports being submitted for the same URL).

Reports containing AI-generated content are low as a proportion of total IWF reports. This may reflect a failure of generative AI to break into the mainstream, or into 'mainstream' CSAM. (Of course, it is not impossible that photorealistic AI-generated content has been missed by analysts in the course of processing reports, and so would not show up in the statistics.)

It is rare, however, for a single topic like generative AI to comprise a large proportion of total IWF reports (notable exceptions are 'self-generated content' and 'ICAP sites', which have different reasons for their high prevalence). In this context, the fact that they comprise a small part of total reports is unsurprising.

What report statistics do not reflect is that where AI CSAM is found, it is more likely than almost all other types of content to be found shared in bulk quantities – large batches of images generated at once. An analysis of bulk-shared AI CSAM images is included later in this section.

The IWF Hotline has limited resources and must prioritise those resources in its fight against the huge amount of online CSAM available. In regard to proactive searching for CSAM, this generally entails focusing on those areas where huge amounts of real child sexual abuse images and videos are known to be found.

It is possible that, in the future, the IWF Hotline focuses more proactive attention on finding online AI CSAM in an effort to increase the statistics provided above. This would mean, however, prioritising searching in places where AI CSAM is likely to be found over the places where real CSAM is known to be found. In addition, as described in section 5, assessment of AI CSAM can be difficult in terms of judging whether content meets the pseudo-photograph criteria for assessment under the Protection of Children Act 1978 – and IWF has generally focused its proactive efforts on this category of content rather than on child prohibited (non-photographic) content.

SEE RECOMMENDATION #2

For Technology Companies

None of this precludes the possibility of a future focus on AI CSAM, nor the possibility that AI CSAM becomes so widespread that it leaks further into 'mainstream' IWF Hotline work, and so is reflected in greater numbers in the statistics anyway. Nonetheless, these facts provide context for the relatively small report figures provided in this section.

Open web and social media

So far, the IWF has found that instances of open web AI CSAM generally follow expected patterns of open web CSAM: realistic pseudo-photographs in areas where real CSAM may be expected, and unrealistic, NPI-style imagery in areas where prohibited images of children may be expected.

The hosting countries in which the 15 open web reports containing AI CSAM were found are provided in the graph below.

AI CSAM hosting countries, May - Oct 2023

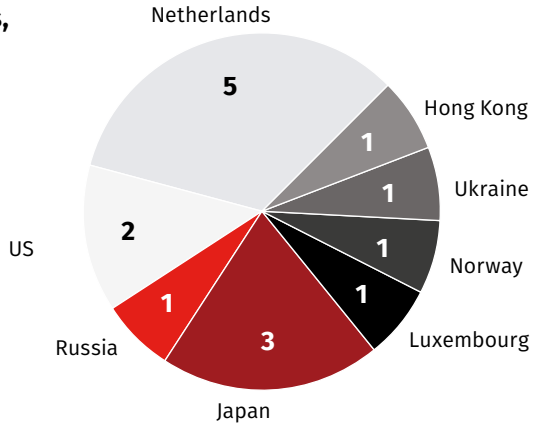


Figure 8
AI CSAM hosting countries

Source: IWF Analysis

Dark web forums

Much of the research for this report, including the majority of quotations used in sections 7 and 9 – verbatim comments from AI CSAM perpetrators – was conducted on dark web CSAM forums.

It is worth emphasising that AI CSAM comprises a small part of dark web forums. The vast majority of these forums are filled with real CSAM (though some are discussion-only, and prohibit the posting of CSAM).

Nonetheless, a number of long-standing forums have this year added new AI sections to their sites – and these sections are growing in popularity.

In these areas, perpetrators share advice on generating AI CSAM; request bespoke images or models; share their work and offer feedback on others' work. These discussions take place openly as users feel anonymous, believing that law enforcement is unwilling or unable to discover them. Users share advice on connecting to these forums in secure, untraceable ways.

Where users share AI CSAM images in bulk on these forums, these files are sometimes hosted on the dark web, and at other times hosted on the open web according to a pre-approved list of secure, anonymous clear web file hosts.

Forum snapshot study: September 2023

This report includes a snapshot study of one dark web CSAM forum. All the AI-generated images posted to the forum in a one-month period (September 2023) were identified.

This encompasses both threads made in September (threads just about AI CSAM, for example, total 261,920 views) and threads made earlier but to which further images were posted in September.

In total, **20,254 AI-generated images were found to have been posted to this forum in a one-month period**.

Of these, **11,108 images were selected for assessment by IWF analysts**. These were the images that were judged most likely to be criminal.

(The remaining 9,146 AI-generated images either did not contain children or contained children but were clearly non-criminal in nature.)

12 IWF analysts dedicated a combined total of 87.5 hours to assessing these 11,108 AI-generated images.

Any images assessed as criminal were criminal under one of two UK laws, as described in section 5. These are:

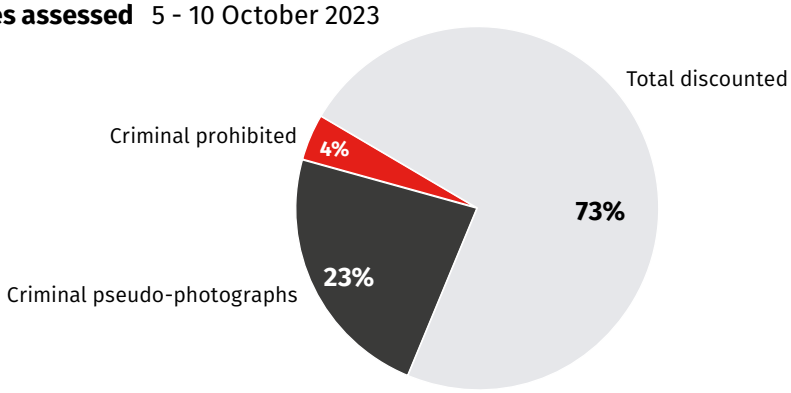
- **The Protection of Children Act 1978** (as amended by the Criminal Justice and Public Order Act 1994). This law criminalises the taking, distribution and possession of an "indecent photograph or pseudo-photograph of a child".
- The **Coroners and Justice Act 2009.** This law criminalises the possession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

2,562 images were assessed as criminal pseudo photographs, and 416 assessed as criminal prohibited images.

Figure 9
AI images assessed between 5 and 10 October 2023

Source: IWF Analysis

AI images assessed 5 - 10 October 2023



These are shown as a proportion of the 11,108 images assessed by IWF analysts as follows:

The total proportion of images assessed as criminal was 27% of the 11,108 images assessed.

Of the criminal images, **six times as many images were assessed as realistic pseudo-photographs than were assessed as non-realistic prohibited images.**

Those images assessed as criminal pseudo-photographs can be sorted by severity (using Sentencing Advisory Panel categories A, B and C) and age:

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.

Figure 10
AI images assessed by severity between 5 and 10 October 2023

Source: IWF Analysis

AI images assessed by severity 5 - 10 October 2023

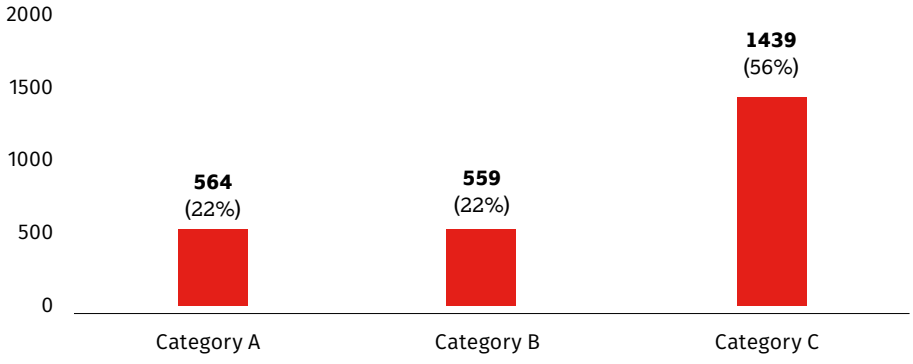
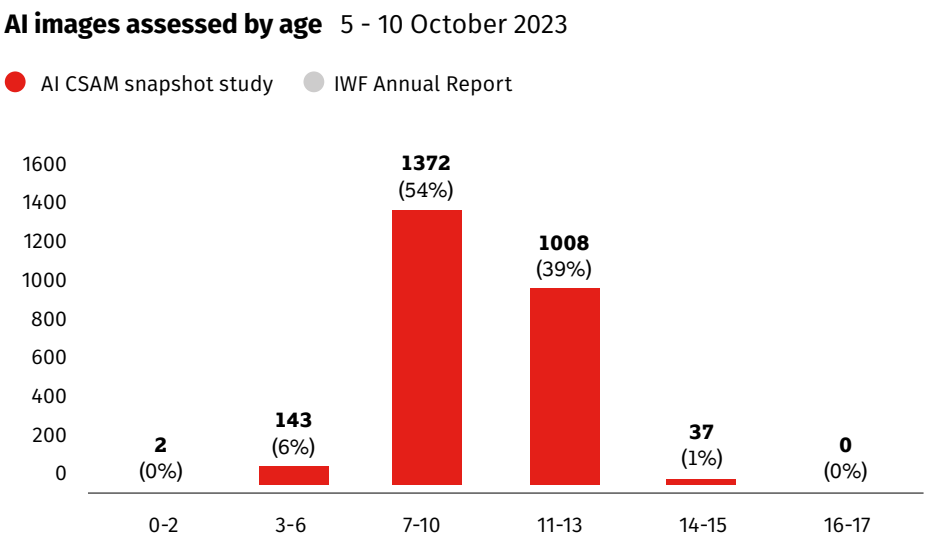


Figure 11
AI images assessed by age between 5 and 10 October 2023

Source: IWF Analysis



These statistics show that **most indecent pseudo-photographs found on this forum were category C** – indecent images of children not falling within categories A or B – often, for example, images depicting naked children erotically posing.

These images were most likely to feature **children between 7 and 13 years old.**

99.6% of these images featured female children. A variety of ethnicities were observed also.

Those images assessed as prohibited images are not sorted by severity, age, or sex. This reflects the limitations of IWF's remit.

The 73% of the 11,108 images that were assessed by IWF analysts as non-criminal can be sorted as follows:

Discounted reason		**%**
None	0	0%
Age in Question	390	5%
Suspected Adult	0	0%
Known Adult	0	0%
Extreme Adult Porn	0	0%
Child No Sexual Activity	4340	53%
Adult No Sexual Activity	492	6%
Non-Photographic Imagery	1634	20%
Off Remit	1274	16%
Total	**8130**	**100%**

The data show that only **20% of assessed discounted images were determined to be not realistic enough** to treat as a pseudo-photograph or a non-photographic image of a child as defined by the Coroners and Justice Act (2009) (whether that AI-generated image depicts a child or an adult). This is roughly in accordance with the proportion of criminal image assessments (between indecent pseudo-photographs and child prohibited images) above.

The 5% of assessed discounted images marked as "Age in Question" reflect cases where the IWF analyst is unsure whether the image depicts a child or an adult – images of older, post-pubescent teenagers, for example. Outside the CSAM world, perpetrators often push the boundaries of generation of adult pornography, behaviour that reflects the wider landscape of online pornography (with its common categories "teens" and "barely legal"). Of course, there is less reason for users of CSAM dark web forums to want to produce these borderline generations, and this is reflected in the low (5%) proportion of "Age in Question" images.

It is notable also that **most assessed discounted images (53%) were images of AI-generated children.** As far as law enforcement is concerned, most or almost all of these would **likely fall into category 6** ("indicative/borderline/notable images") – nudist, naked or semi-naked images of children that have legitimate settings or do not meet the threshold for indecency for their assessment as criminal. These images, however, do not meet the threshold for IWF to take action against them.

This snapshot study had necessary limitations:

- 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 number of regular 'creators', and so large batches of images assessed originate from the same few perpetrators.

Further study would test whether the key findings of this snapshot study would be replicated in, for example, a forum that leaned towards images of boys, or 'hardcore' CSAM.

Nonetheless, these key findings are summarised as follows:

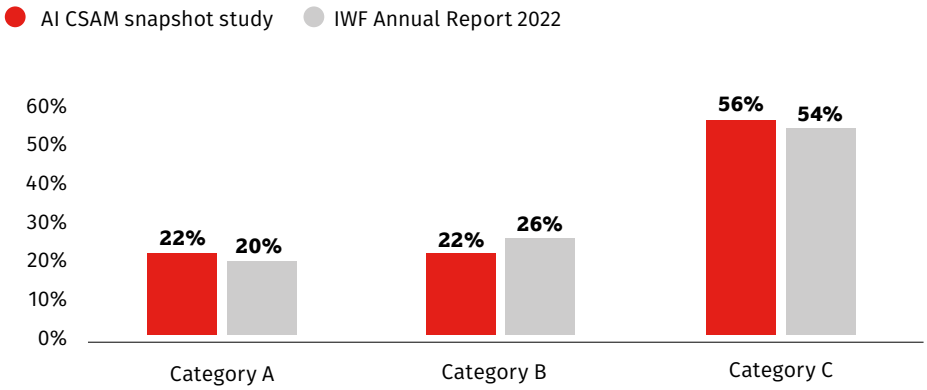
- Most **AI-generated images assessed were realistic enough** to meet the realism threshold for assessment as a pseudo-photograph of a child (if criminal). This finding holds across both assessed criminal and assessed non-criminal images. The high level of realism seen among images found on the forum is owing to a number of factors, as discussed in section 5.
- Most **AI-generated images assessed were not criminal.** This reflects a large appetite for images of children outside scenarios containing explicit sexual activity (Sentencing Advisory Panel categories A-C).

- AI CSAM was found **to reflect the wider CSAM landscape.** Some analyses of AI CSAM suggest that AI imagery tends towards the most extreme categories of content, and the youngest ages of children. While this may be true for some areas of the internet, this study found that category and age assessments of criminal AI-generated pseudo-photographs of children were closely correlated with assessments of CSAM across the internet as a whole. This can be most clearly demonstrated by comparing the category assessments from this snapshot study to the category assessments reported by the IWF last year for content found on the internet as a whole:

Figure 12
Severity of assessed criminal content

Source: IWF Analysis

Severity of assessed criminal content



9

AI CSAM: issues

Perpetrators can turn to AI CSAM over real CSAM for many reasons. One dark web forum user provides a typical view in listing five benefits:

> *1. Users can create images on their own devices;*
>
> *2. Everything can be custom-made and edited to specification;*
>
> *3. AI CSAM can show what is impossible in the real world;*
>
> *4. It is very secure;*
>
> *5. AI is always getting better.*

To some extent, all these points are true.

Realism

The level of realism of AI CSAM varies between individuals who generate the content and between image sets. Technical expertise, computer size, and time invested are all variables that affect the level of realism.

Images of simpler composition, generally showing just one (child) character, are more likely to look photorealistic – there is a greater likelihood of AI artefacts appearing in images with multiple characters involved in complex activity. Nonetheless, the abundance of post-generation editing tools available means that this aspect can be overcome with enough technical knowledge and time investment.

For all these reasons, contrasting reactions can be found across forums where AI CSAM is shared. Some are dismissive:

> *"I can recognize AI work at a glance"*

Others express surprise or admiration at the quality of the output:

> *"It's been a few months since I've checked boy AI. My God it's gotten really good!"*

> *"These are truly stunning. Some of the realism in these is about 95% of the way to indistinguishable from real photos."*

> *"How the hell can you get this kind of images? I've seen realistic images but this is superb."*

> *"The AI generated images are getting better and better."*

> *"The photorealism here is stunning, I mean I'm sure a trained eye can still see it's a generated image, but not by much."*

Others question whether they are looking at real images, or claim photorealistic AI-generated images:

> *"I just can't get my head around that these boys are not real!"*

> *"Are you sure these are CGI?"*

> *"Congratulations on hitting photo realism. This is insane. How much better does it get from here?"*

Examples of individuals asking, 'who is this?' in reply to images – only to be informed that the images are not of real children – have been found.

Forum users' exposure to CSAM and to AI CSAM varies, as does their inclination (or disinclination) towards both. IWF analysts, on the other hand, are exposed to CSAM every day. They are trained to recognise and assess CSAM, and in recent months have been trained also to assess AI CSAM.

For this report, IWF analysts assessed thousands of AI-generated images, and provided comments on what they thought of the level of realism. Their comments provide a useful perspective on comparing AI CSAM to real CSAM; on assessment of AI CSAM; and on the future outlook for those who work to fight child sexual abuse, including law enforcement. Some have stronger conclusions on realism than others.

ANALYST 1

"We have a good idea of the common glitches and features of AI-generated images. Armed with that knowledge and assessing images that I know are AI-generated, there are still images that I would struggle to distinguish from real photos. Near flawless, photo-realistic pictures of the worst kind of child abuse you can image. And this is with AI in its infancy."

ANALYST 2

"Currently AI generated images are quite simple to spot as the tells such as extra fingers, lighting, etc. are still quite prominent, however I think that the quality has improved very quickly over a short period and would confuse the general public."

ANALYST 3

"I've been both surprised and disappointed to see how much attention and dedication has been taken to create such life-like abusive images of children."

ANALYST 4

"Some were scarily realistic, and the whole thing just made me feel a bit uneasy to be honest... I am also concerned that future images may be of such good quality that we won't even notice."

Further thoughts from analysts on making assessments of AI CSAM are included in section 11.

Satisfaction

How realistic AI CSAM looks is closely associated with discussions among perpetrators about whether AI CSAM satisfies their desires.

In some areas, AI CSAM is shared alongside real CSAM (a 'mixed' environment). In these areas, the preference is for realistic imagery:

> *"less they look cartoon the better."*

> *"Please more of the Ultra Realistic Stuff. Amazing."*

> *"Really cartoonish or unrealistic AI images don't really do it for me, but they are getting better. The image in this post is excellent!"*

Nonetheless, there remains parts of CSAM communities (indeed, perhaps the vast majority of CSAM communities) to whom AI CSAM is not interesting, or does not match real CSAM. These individuals may only be attracted to real abuse.

> *"AI is completely uninteresting to me."*

> *"...it's nothing to the real thing."*

Some users abuse others for posting AI-generated images. One AI CSAM perpetrator says:

> *"I don't get the abuse, I like to make these and ok they are not perfect but they are nice and fun and look good."*

Whether such opinions hold where AI CSAM is not photorealistic, or whether they hold in principle – no matter the appearance of AI-generated images – is questionable. A stronger anti-AI CSAM opinion holds:

> *"They will be saved by people and dilute the stock of real pictures. Do this for five years and fakes is all we will have left! AND THEY WILL STILL LOCK YOU UP FOR THEM."*

Some users claim that AI CSAM will never match real CSAM because it lacks the

> *"sense of danger."*

Other AI CSAM perpetrators disagree. One comment even claimed that AI CSAM was superior because real CSAM images are often low-quality:

> *"Most pics don't come anywhere close to the quality produced by AI. Poorly composed, poorly exposed, low res, and out of focus pics are common."*

Another user wanted to produce lower-quality AI-generated images in order to make them more like real CSAM:

> *"By default AIs weren't teach that reality is sick & dirty, because people want beautiful and perfect pictures so they instruct them that way. But when it comes to the sexy, something is missing: sick & dirty is part of it! That's why I'm not satisfied and try to find ways to get more realistic rendering."*

Because AI-generated images have by default a high-quality, 'clean' appearance, images produced according to this kind of opinion – images that appear lower-quality, 'grainier' or less clear – may pose a key challenge for future AI detection efforts.

Ethicality and legality

Many individuals claim that AI CSAM is more ethical than real CSAM and use this claim as a justification for generating and posting AI CSAM.

> *"the future of CP is already here... and not offending anyone..."*

Others emphasise that their AI CSAM images were generated without CSAM fine-tuned models, perhaps as an effort to legitimise those generations:

> *"All of these were created without any real world child porn whatsoever."*

At an extreme end, some perpetrators claim that AI-generated images comprise the future of CSAM – eventually replacing the need for real CSAM:

> *"[AI CSAM] makes CSAM unthinkable. Anyone who might before have justified needing CSAM in order to quell some irresistible urge will have no more excuses."*

Representative of such claims is the widespread use of disclaimers among perpetrators posting AI CSAM both on the clear web and dark web:

> *"All images are A.I. generated and none existed before I entered the very specific text prompts I entered to create it"*

> *"Disclaimer: none of the boys I'll post in this thread are real, they are all generated by an AI"*

The reason for these disclaimers is unclear, and may vary between perpetrators. They may intend these disclaimers to discourage takedown efforts from site owners or investigation by law enforcement, or they may be for information purposes only.

Users discussed how law enforcement agency (LEA) officers could generate and use AI CSAM, and how they could use AI against LEAs:

> *"For now AI images can be spotted in most cases, but it is getting to the point where AI will be indistinguishable. I think there are opportunities for AI to be used to our advantage in playing an activist role."*

As the latter comment suggests, some users discuss sharing AI-generated images with non-perpetrators as an intended 'gateway' to real CSAM.

Perpetrator pathways

Reflecting IWF's remit, concrete knowledge of pathways of perpetration or offending is elusive, but anecdotal evidence encountered during research for this report suggests that movement between AI-generated non-photographic images and AI-generated pseudo-photographic images is possible, and, therefore, between AI CSAM and real CSAM in principle.

> *"I've mostly been using AI to generate smut (haven't we all) and recently moved from semi-realistic [cartoon-style] content, which I had some great successes with, to trying to generate photoreal children."*

Worth re-emphasising in this section is that though some perpetrators use complex models and fine-tuned CSAM models, for others, simple, easy-to-use models exist and can be used to generate AI CSAM. One perpetrator describes:

> *"A picture of a couple random kiddies I seen playing outside at the park one day I went and made it a fake"*

The prevalence of tools like those for 'nudifying' images makes such low-effort perpetration possible. Recent news reports detailed the use of this technology by children, applied to images of children, among students at a school in Spain and another school in Denmark.

Figure 13
Recent BBC News article on 'nudification' technology spread in a school in Spain

Source: BBC News
Author's screenshot

The IWF has already seen some examples of self-reporters claiming that their images had been turned into sexual images with AI (possibly using these 'nudifying' tools), and is aware of a number of websites that are claimed to have been used for this purpose.

Clearly, the ease and accessibility of some text-to-image technology increases the risk among the pool of potential perpetrators – a major area of concern both in terms of isolated incidents and in terms of targeted 'sextortion'.

Guides to generating AI CSAM

Section 7 described the two main complete guides so far discovered circulating on dark web forums: a guide for creating CSAM images, identified in March 2023 (such is the pace of technological progress, whose contents should now be considered outdated); and a guide for creating new models, identified in May 2023. Other small or one-page guides concerning, for example, online web-based models have also been found.

SEE RECOMMENDATION #2

For Government

Just a few guides have so far been identified, though it is possible that some are in circulation of which IWF is not aware.

Why are there not more guides in circulation? A huge amount of information on generating AI CSAM is shared, but not as cohesive guides. Instead, these are individual posts on forums. Of course, a high amount of effort is required to write, edit, and share these documents, so perhaps their low frequency is unsurprising.

Whether more guides appear in the coming months remains to be seen. Perhaps the instability caused by the current pace of technological progress in this area is incompatible with production of long, detailed guides on using text-to-image AI, and more guides will appear when the technology stabilises. Alternatively, the time and effort barrier to the production of guides – which have a highly technical subject matter – may mean that not very many guides will ever appear.

Nonetheless, guides do exist – and their legality is discussed in section 11.

Commerciality

The first examples of commercialisation of AI CSAM have been recorded over the past months. Examples are limited because the IWF is often prevented from further investigation by direction to end-to-end encrypted chats or peer-to-peer networks; the IWF is also unable to purchase, or attempt to purchase, AI CSAM. Intelligence collected by law enforcement partners in publicly inaccessible areas would support examples provided in this section.

- **Commercial example 1**

An example of the barriers IWF face can be found in a page called [redacted]. The account posted non-actionable AI-generated images of children – in various modelling poses, settings, and outfits – and linked to a website and another page through which people could purchase access to more content. The brand claims that it is

```
selling custom AI images
```

but the authenticity of this claim cannot be verified. Interested individuals have to

```
hit up my [name] on secret chat
```

to access this hidden content. Whether hidden content comprises just further child modelling images or also AI CSAM is impossible for IWF to know at present.

- **Commercial example 2**

Another commercial page showcased non-actionable images and GIFs of AI-generated children (including clothed; posed suggestively; in bikinis), with further content locked behind a paywall. Previews showed high-quality AI-generated content apparently produced by:

> Harnessing the power of the latest AI technology and [a high-end computer graphics card]

Supporters can apparently view more than 2,000 images and request 10 new images monthly for the price of 5,000¥ per month. Again, no criminal content was accessible and so whether only non-actionable images are found behind the paywall, or whether AI CSAM is being sold and bought, is impossible for IWF to know at present.

- **Commercial example 3**

This commercial example is associated with a website actioned by IWF as containing criminal preview AI CSAM material.

An account advertised

> NSFW ai generated pictures

including bespoke content, generated to specification. Individuals could subscribe from between £4.50/month to £17/month. At the time of assessment, the account had 70 total paid members, generating $316.50/month for the page operator. (This account has since been closed.)

- **Commercial example 4**

A wide-ranging commercial brand, again linked to actioned preview AI CSAM content offered customers:

> Get access to exclusive, photo-realistic arts (1000+ for now). With regular updates and unique personal models. (requests accepted)

The access price was $7/month and promised more content of naked AI-generated children behind this paywall.

These four examples show the demand for high-quality AI-generated images of children and AI CSAM made to specification. They demonstrate the growth of brands that advertise this service and are careful about hiding most or all criminal content behind a payment barrier. It is also notable that these services are international – perpetrators may reside in the UK or in other territories in which AI CSAM has a different legal status.

SEE RECOMMENDATION #1

For Government

Detection and enforcement

AI image detection

The goal of detecting AI-generated media has broad relevance – not just in the CSAM domain, but also for law enforcement more broadly, implications for identifying cases of misinformation and disinformation, and democracy as a whole.

There is growing pressure for standards for digital watermarking of AI-generated images; there is also an increasing industry for accurate AI image detection. Tools can take the form of online services or downloadable software.

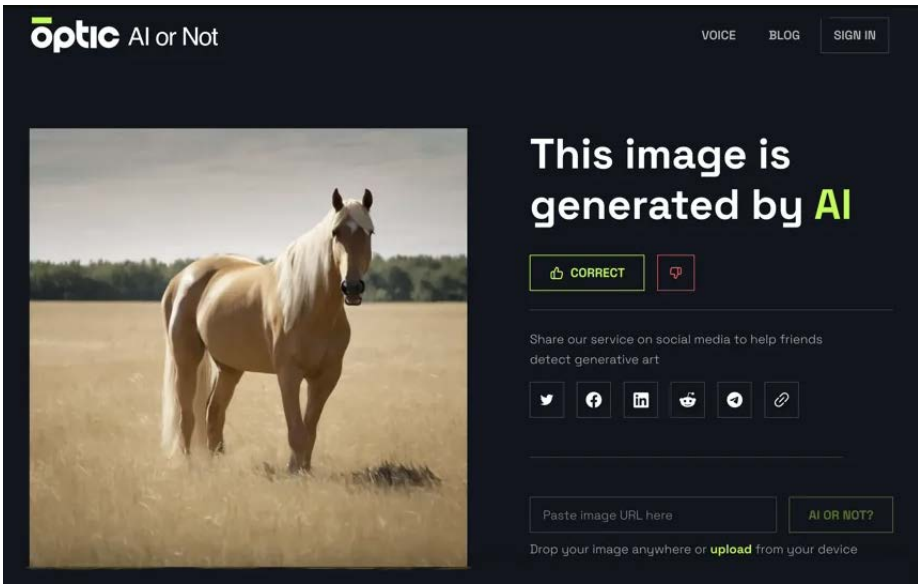


Figure 14
Optic's website 'AI or not' aims to answer this question for each image uploaded to the site

Source: PCMag

In the CSAM domain, US nonprofit Thorn have produced an open-source detection tool, and claim high accuracy.

Concerns exist that a race is beginning between AI classifiers and those training text-to-image models to evade AI classifiers. This represents a major barrier to the development of a classifier that is 100% effective.

This race is possible because a favoured model being used to create AI CSAM is open-source technology. Data that comprises an AI watermark can be added to real images; the watermark can be removed from AI-generated images.

The first example of a perpetrator claiming that real images were AI-generated has been identified by law enforcement; the provenance of these images was discovered through a check against the UK's national Child Abuse Image Database (CAID).

Reasons for this type of perpetrator activity are mostly speculative – only anecdotal evidence exists. One discussion found between users of a dark web forum concerned whether or not perpetrators should 'sign' their own AI CSAM (as an 'AI artist'), and whether this would help law enforcement agencies (LEAs). At the very least, this indicates that perpetrators are aware of some of the difficulties for law enforcement:

> *"I disagree, by signing your work you make LE job easier. If they know an author is creating AI images then they can ignore it. If an image looks real they have to spend time 'looking' into it, which wastes their resources."*

Of course, a classifier does not need to be 100% effective to be useful to IWF analysts or to law enforcement. A more moderate goal revolves around provision of a probabilistic tag-and-flag tool, allowing law enforcement, for example, to prioritise cases in which it is more likely that there is a child that needs protecting. Further discussion about AI CSAM and LEA is found later in this section.

SEE RECOMMENDATION #8

For Technology Companies

The current prevailing view seems to be that future AI image detection will require a suite of classifiers, perhaps targeted at detecting the output of different text-to-image models, whose analyses will in conjunction indicate the likelihood of an image having been AI-generated.

Model types and model data

As outlined in section 4, a popular base model is trained on subsets of an open-source dataset, which is an index of URLs that identifies billions of images. Earlier versions of this base model contain adult pornography, and base models after a certain point exclude adult pornography. This does not completely prevent generation of adult pornography with the later models but does make users more likely to use earlier models for the generation of adult pornography – and for the generation of CSAM.

In this context, the company behind this base model must perform some filtering within the dataset to create various versions of the base model. How exactly it performs this filtering – whether manual, automatic, or a combination of both – is unknown.

The key question is: given a diffusion model, what (if anything) can be known about the training dataset?

These questions comprise an active area of study among the AI research community. For example, a 2023 paper asked whether training data could be extracted from diffusion models by generating images, then performing membership inference to identify 'memorised' images. On model memorisation, the authors note:

> ***"This paper covers a very restricted definition of "memorization": whether diffusion models can be induced to generate near-copies of some training examples when prompted with appropriate instructions. We will describe an approach that can generate images that are close approximations of some training images (especially images that are frequently represented in the training dataset through duplication or other means). There is active discussion within the technical and legal communities about whether the presence of this type of "memorization" suggests that generative neural networks "contain" their training data."***

Researchers found significant rates of memorisation in large diffusion models, a finding that could be used in a limited way to 'reconstruct' images within a training dataset by testing and evaluating the output of diffusion models.

Such methods apply to large-scale diffusion models, but what about smaller fine-tuned models? Could the training dataset for a CSAM fine-tuned model be reconstructed, or any information at all be deduced, from the CSAM fine-tuned model only?

These questions require further study.

AI CSAM, victim identification, and law enforcement

As mentioned earlier in this section, the major challenge for law enforcement posed by AI CSAM is of distinguishing photorealistic AI CSAM from real CSAM – victim identification (VID). This challenge is to be addressed through technological solutions – tools like AI classifiers, as described above – and advanced digital forensic knowledge among investigators.

SEE RECOMMENDATION #4

For Law Enforcement and Regulators

LEAs will increasingly be required to train their investigators on recognising AI-generated images; investigating, assessing, and tagging them appropriately (considering, for example, the impact of uploading AI-generated images to the UK's Child Abuse Image Database (CAID)).

There is a low risk for IWF of false positive victim identification referrals – referring virtual children to law enforcement for investigation. This is because IWF analysts require a baseline amount of identifying information to make a referral – the sort of identifying information (name; school; region or area; or similar) that would be expected to be missing from AI-generated images.

Nonetheless, because image generation is fast and accessible (especially for low-tech perpetrators), as the examples in section 9 of this report show, this technology increases the pool of potential victims of child sexual abuse.

As stated in the introduction to this report, generative AI more broadly has the potential for misuse in CSAM and CSE/A offending. This includes misuse of LLMs and chatbots, and use of text-to-image technology to generate child avatars, for example, or other images that increase a perpetrator's repute

among potential victims. These uses merit further investigation but fall somewhat out of the scope of this report.

SEE RECOMMENDATION #4

For Law Enforcement and Regulators

The speed and scale of potential AI CSAM generation should concern LEAs – faced, perhaps, with investigating seized devices that contain vast amounts of AI CSAM generated offline, on-device – as well as IWF – faced with the potential for the spread of vast amounts of AI CSAM across the internet.

SEE RECOMMENDATION #5

For Law Enforcement and Regulators

Finally, the lack of oversight inherent in the open-source technology that is overwhelmingly favoured by AI CSAM perpetrators should concern LEAs. There are few areas for oversight, detection, or intervention – and a clearly conceivable route of offending from realistic AI-generated CSAM to real CSAM.

11

UK legislation

Assessments

The key feature of AI CSAM from an assessment perspective is that, as described in section 5, it straddles two pieces of legislation. This entails an often-difficult decision for those responsible for classifying images: whether an image is realistic enough to classify as an indecent pseudo-photograph, or whether it should be assessed as a prohibited image of a child. (The latter piece of legislation has more criteria and carries a shorter sentence for convicted offenders. The question of what position a judge would take on a conviction on AI-generated indecent pseudo-photographs of children remains to be tested in UK courts. A lesser sentence is a potentiality, as a judge may conclude that no harm has been done to an actual child.)

Assessment difficulties were widely reported by IWF analysts, who assessed thousands of AI-generated images for this report and hashed them – translated those images into code that can be used to identify and remove those images in the future. Such difficulties entail significant time investment. Some of these comments are reported below.

ANALYST 3

"It can feel odd to question the realness of something that you know isn't real. It brings into focus the different laws our work is bound by."

ANALYST 5

"Hashing these images were difficult, more difficult than 'normal' CSAM. Is it a child? Is it actionable? Is it photorealistic? Is the actionable content you are seeing feasible even?"

"There were a lot of digitally anatomical disturbing images of all sorts which take a lot longer to work out and grade."

ANALYST 6

"There are some very weird concoctions which are more difficult to grade, and other opinions are sought to help with them."

An argument could be made that the emergence of AI CSAM combined with the difficulty of sorting between laws and categories – indeed, depending on individual assessors' determinations of whether an image "appears to be a photograph" – merits combining existing CSAM legislation.

SEE RECOMMENDATION #2

For Government

Certainly, this would dissolve the difficulties of assessing AI CSAM, but it would create new difficulties.

International mapping remains a problem. As discussed in section 8, AI CSAM has different legal statuses in different jurisdictions.

Recommending specific change(s) to this existing legislation is beyond the scope of this report, but a future review in this area, including input from law enforcement partners, should inform any potential recommendations.

SEE RECOMMENDATION #2

For Government

As of September 2023, AI-generated imagery is not yet part of the College of Policing grading training course – organisations working in this area must work to come to terms with the problem fully before solutions are considered.

RECOMMENDATION #4

For Law Enforcement and Regulators

Guides

Section 9 of this report briefly set out the phenomenon of guides to generating AI CSAM, shared in dark web forums, and covering topics like generating and editing images, and training CSAM fine-tuned models.

Serious Crime Act 2015 created the following offence as part of Section 69, "Possession of paedophile manual":

> It is an offence to be in possession of any item that contains advice or guidance about abusing children sexually.

But "abusing children sexually" means doing anything that constitutes an offence under Part 1 of the Sexual Offences Act 2003 or:

> (b) an offence under section 1 of the Protection of Children Act 1978, or under Article 3 of the Protection of Children (Northern Ireland) Order 1978, involving indecent photographs (but not pseudo-photographs).

The result is that guides to the generation of AI CSAM are not covered by Section 69 of the Serious Crime Act 2015.

Nonetheless, there is an argument that this falls into the 'low priority but easy to solve' category, wherein guides to generating or 'creating' indecent pseudo-photographs of children could be added to the definition set out in the legislation above.

SEE RECOMMENDATION #2

For Government

Models

Articulation of an offence that criminalises CSAM fine-tuned models is difficult, and it could be argued that possessing (or even creating) such models comprises only a preparatory act, not a criminal one. Furthermore, technical questions about how to prove that a given model has been fine-tuned using a CSAM dataset, and is intended for the generation of AI CSAM, remain.

Not to criminalise these models risks their widespread distribution across CSAM communities – at present, any perpetrator can download everything that they need to generate (offline, undetected) as many images of known victims of child sexual abuse as they without committing any offence. Clearly, the other items have widespread legitimate uses, but it is difficult to argue that the final item does have any legitimate use.

As such, it is unclear whether the IWF nor any regulator or law enforcement body currently has recourse to request removal of CSAM fine-tuned models from legitimate model-sharing sites, for example.

Two possible routes exist: firstly, where the model has been shared alongside criminal preview images; secondly, and more tentatively, where provision of the model can be judged to constitute an offence under the Serious Crime Act 2007 ("encouraging" an offence to be committed).

SEE RECOMMENDATION #7

For Technology Companies

This report makes no conclusion on whether it is desirable or even possible to reconcile the contradictory positions on this issue.

Summary: past, present, and future of AI CSAM

Progress in computer technologies, including progress in generative AI, has enormous potential to better our lives, and misuse of this technology is a small part of this picture.

The development of computer technologies like the growth of the internet, the spread of video-calling and livestreaming, and the development of CGI and image-editing programs, have enabled the widespread production and distribution of CSAM that is currently in evidence.

It is too early to know whether generative AI should be added to the list above as a notable technology that comprises a step change in the history of the production and distribution of CSAM.

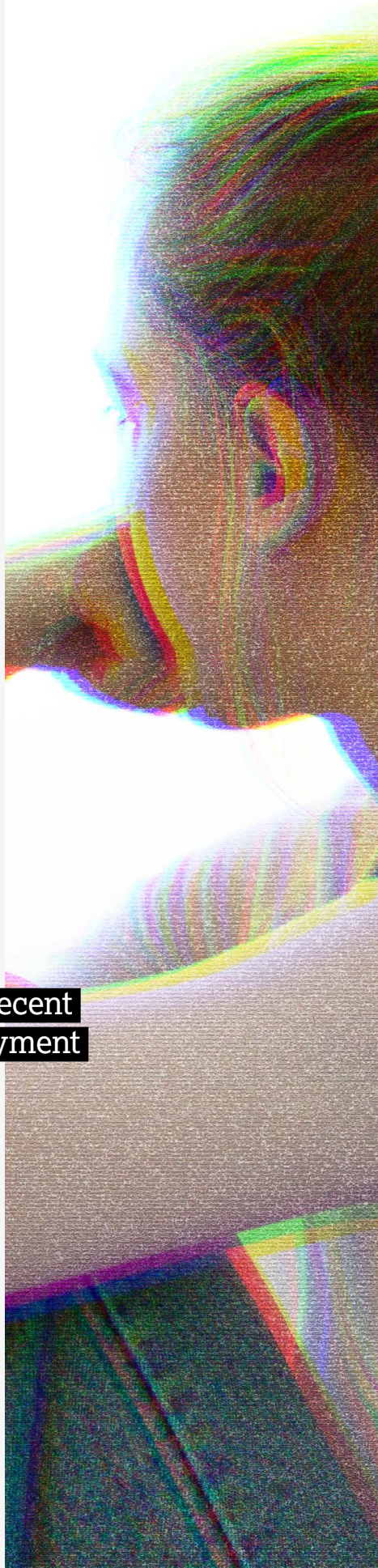
Nonetheless, this report evidences a growing problem that boasts several key differences from previous technologies. Chief among those differences is the potential for offline generation of images at scale – with the clear potential to overwhelm those working to fight online child sexual abuse and divert significant resources from real CSAM towards AI CSAM.

In this context, it is worth re-emphasising that this is the worst, in terms of image quality, that AI technology will ever be. Generative AI only surfaced in the public consciousness in the past year; a consideration of what it will look like in another year – or, indeed, five years – should give pause.

At some point on this timeline, **realistic full-motion video content will become commonplace. The first examples of short AI CSAM videos have already been seen – these are only going to get more realistic and more widespread.**

Solving some of the problems posed by AI-generated indecent images now will be necessary to create models for deployment against the growth of video content in the future.

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



Glossary

Actionable *(image)*: an actionable image is one that is deemed criminal under UK law and therefore IWF can seek its removal from the internet.

AGI: *artificial general intelligence.*

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.

ChatGPT: An LLM developed by OpenAI.

Claude: An LLM developed by Anthropic.

CLIP: *contrastive language-image pre-training.* A neural network trained on hundreds of millions of text/image pairs scraped from the internet.

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.*

C2PA: A mode of metadata representation used for digital watermarking of AI-generated images.

DALL-E: A text-to-image model developed by OpenAI, accessed through an API.

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.

Deep learning: A type of machine learning, loosely modelled on the human brain, that uses artificial neural networks with more than three layers. These deep learning systems are generally trained on huge datasets scraped from the internet.

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.

Fine-tuning: A type of machine learning model in which the weights of a pre-trained model are trained on new data, and therefore adjusted, to perform a secondary task.

GANs: *generative adversarial networks.* A type of machine learning model in which two neural networks compete with each other by using deep learning methods to become more accurate in their predictions. Can be considered the precursor to diffusion models.

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.

ICAP sites: *invite child abuse pyramid sites.* Sites first reported on by IWF in June 2022 that incentivise users to share links to child sexual abuse webpages far and wide in a 'scattergun' approach.

Iterative learning: a type of machine learning guided by a combination of human feedback (supervised machine learning) and/or algorithmic feedback (unsupervised machine learning).

IPTC: A mode of metadata representation used for digital watermarking of AI-generated images.

IWF: *Internet Watch Foundation.*

LAION: An open-source dataset used for training Stable Diffusion which has 2.6 billion English language-tagged images within a 6 billion image dataset.

LEAs: *law enforcement agencies.*

LLaMA: An LLM developed by Meta.

LLMs: *Large Language Models.* A type of machine learning that is trained on huge quantities of text and whose function is to generate text. These models are renowned for their natural language processing abilities, 'understanding' and interpreting human language.

Midjourney: A text-to-image model developed by Midjourney, Inc, accessed through the social media site Discord.

Neural network: a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain. It creates an adaptive system that computers use to learn from their mistakes and improve continuously.

NSFW: *not safe for work.*

NPI: *Neural Programmer-Interpreters.* A machine learning model that uses a recurrent and compositional neural network to train machines to carry out simple tasks based on a small amount of training data.

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 Web: The side of the web that is public and viewable by everyone.

PaLM: An LLM developed by Google.

PoC: *The Protection of Children Act 1978.* This law criminalises the taking, distribution and possession of an "indecent photograph or pseudo-photograph of a child" (as amended by the Criminal Justice and Public Order Act 1994).

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.

SDXL: A new version (released July 2023) of Stable Diffusion, the text-to-image model developed by Stability AI.

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

Serious Crime Act 2015, Section 69. This section created the offence of "possession of a paedophile manual".

Shallow Fake: Colloquial term encompassing images produced via simple image editing tools or software, in contrast to deepfakes, which generally use AI tools or software.

Stability AI: a global company, headquartered in London, working to make foundational AI technology accessible to all. Responsible for the creation of Stable Diffusion.

Stable Diffusion: A text-to-image model developed by Stability AI that can be downloaded from an open-source online community.

SynthID: A Google technology that adds a digital watermark to individual pixels in AI-generated images.

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

VAEs: *variational autoencoders.* A generative AI model that uses two neural networks called the encoder and decoder. Generally, these can output images faster than diffusion models, but those images are less detailed.

VID: *victim identification.*

Watermarking/digital watermarking: a technique that involves embedding digital marks or indicators into machine learning models or datasets to enable their identification.

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 someway to demonstrate the photorealism of images produced by AI models.