

GENERATIVE AI

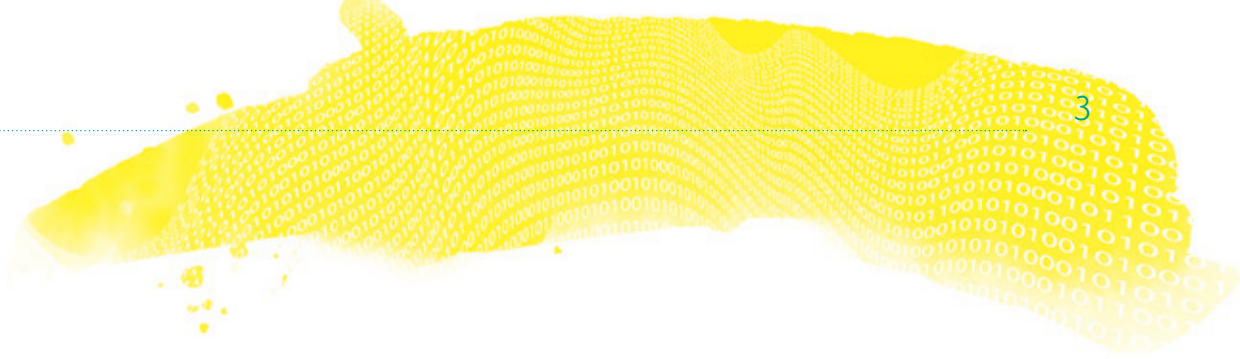
A New Threat for Online Child Sexual Exploitation and Abuse



CONTENTS

EXECUTIVE SUMMARY	4
Scope of this Report	5
Terminology Used in this Report	6
I. “We Cannot Arrest Our Way Out of this Problem”	8
THE EMERGING THREAT OF AI-GENERATED CSAM	
What Is Generative AI?	8
How Is Generative AI Used to Create CSAM?	9
What Types of CSAM Can Generative AI Create?	10
Estimating the Prevalence of CSAM and AI-Generated CSAM	12
How Is AI-Generated CSAM Impacting the Landscape of Threats to Children Online?	15
II. “Enforcement Can Only Move as Fast as the Speed of Law”	23
CURRENT LEGAL STATUS OF AI-GENERATED CSAM	
Legislating AI	23
Legislating AI-Generated CSAM	24
Unique Challenges in Legislating AI-Generated CSAM	26
III. “Even Offenders Won’t Be Able to Differentiate”	31
DETECTING AI-GENERATED CSAM	
Law Enforcement and Detection	31
Tech Companies and Detection	32
The Future of Detection for AI-Generated CSAM	33





IV. “We’re Behind the Curve”	34
HOW ARE PRIVATE COMPANIES, GOVERNMENTS, LAW ENFORCEMENT, AND CAREGIVERS RESPONDING?	
Private Sector	34
Law Enforcement	36
Governments	36
Caregivers	39
V. “The Challenge of the Decade”	42
RECOMMENDATIONS FOR COUNTERING THE THREAT	
ACKNOWLEDGMENTS	44
ABOUT THE AUTHORS	45
FURTHER READING	46
BIBLIOGRAPHY	47
IMPRINT	52

.....
Disclaimer:

This report has been prepared by Bracket Foundation and Value for Good GmbH in cooperation with the Centre for Artificial Intelligence and Robotics at the United Nations Interregional Crime and Justice Research Institute (UNICRI). The opinions, findings, conclusions, and recommendations expressed herein do not necessarily reflect the views of the United Nations or any other national, regional, or global entities involved. The designation employed and material presented in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city, or area of its authorities, or concerning the delimitation of its frontiers or boundaries. Contents of this publication may be quoted or reproduced, provided that the source of information is acknowledged. The authors would like to receive a copy of the document in which this publication is used or quoted.

EXECUTIVE SUMMARY

Online sexual exploitation and abuse is endangering children across the world.¹ In fact, one out of eight children around the world has been the victim of online sexual exploitation.² The production, possession, and distribution of child sexual abuse material (CSAM) is a significant component of online child sexual exploitation and abuse, resulting in long-term harm to victims. CSAM includes any representation of a child engaged in real or simulated explicit sexual activities and already threatens children across the globe.

The rise of artificial intelligence (AI) is altering the landscape of child harm. One of the most pressing dangers facing the global child protection ecosystem is AI's effects on CSAM. Of the approximately 36 million reports of online child exploitation and abuse to the U.S. regulator in 2023, 4,700 included verified AI-generated CSAM.³ While this number is still comparatively low, experts anticipate a stark increase as the capabilities to create AI-generated CSAM advance.⁴ Understanding the landscape of AI-generated CSAM is a necessary precursor to safeguarding children. This study explores the current forms of AI-generated CSAM, which include:

- **Text content:** Generative AI chatbots have been shown to engage in sexually explicit chats acting as children might and generative AI has also generated guides, tutorials, and suggestions on how to sexually abuse children.⁵
- **Still & moving imagery:** Generative AI models are increasingly able to generate photorealistic CSAM and alter existing imagery to make it explicit. As this technology improves, perpetrators can create higher quality moving imagery and videos.

Generative AI has also expanded the ways in which children can become victims. The wide range of victims includes:

- Children whose (innocuous) images have been used to train AI models
- Children whose innocuous images are transformed into CSAM with AI
- Existing victims of CSAM, who have been revictimized through the modification or obscuration of existing CSAM
- Adults whose images have been de-aged to create AI-generated CSAM

In addition to adult perpetrators of child sexual abuse, young people themselves are increasingly becoming creators of AI-generated CSAM through the use of "nudify" apps.

Pulling together perspectives and data from law enforcement, private sector, governments, and caregivers, this report presents a comprehensive view of the main challenges affecting stakeholders.



- **Private sector:** Tech companies, like social media platforms, must contend with the role their platforms play in spreading AI-generated materials and determine how to integrate safety features. Meanwhile, AI developers must balance ethics and technological progress.



- **Governments:** Governments around the world are assessing how best to legislate this threat to children's safety. Policymakers must explore legal methods that balance protecting children while still encouraging technological development.



- **Law enforcement:** Law enforcement is grappling with how to detect AI-generated CSAM and safeguard the children

¹ "Into the Light Index," Childlight Global Child Safety Institute, accessed June 12, 2024.

² Andy Gregory, "Prioritise Children's Online Safety at Election to Tackle 'Hidden Pandemic' of Sexual Abuse, Experts Urge," *The Independent*, June 2, 2024.

³ John Shehan, "Addressing Real Harm Done by Deepfakes," NCMCEC, 2024.

⁴ Shehan, "Addressing Real Harm."

⁵ Shehan, "Addressing Real Harm."

who are potentially harmed by perpetrators. As new criminal business models arise, law enforcement must stay abreast of the latest technological and criminal developments.



- **Caregivers:** Parents, educators, and health professionals must remain aware of the dangers facing children in the context of AI-generated CSAM and endeavor to protect the children in their care.

To address this rapidly evolving threat, these stakeholders must come together to safeguard children and prevent the production, distribution and commercialization of AI-generated CSAM. Key recommendations for addressing this threat are summarized in this report, including:



- **Private sector:** AI developers must implement safety measures to prevent models from generating explicit content, particularly involving children. Tech platforms should prioritize children's safety by blocking and moderating AI-generated CSAM, cutting distribution channels.



- **Governments:** Policymakers must update laws to address AI-generated CSAM, requiring systemic reforms and increased investments in technology. Collaborations with tech providers are essential to ensure robust child safeguards.



- **Law enforcement:** Law enforcement needs to stay updated on AI-generated CSAM trends through international exchanges and adopt new tools for identifying such content, ensuring effective responses.



- **Caregivers:** Caregivers must stay informed about online threats to children, openly discuss internet dangers, and utilize available resources. Limiting children's online presence should also be considered.

Scope of this Report

Reports are growing that AI capabilities are exposing children around the world to a new type of online child sexual exploitation and abuse: AI-generated CSAM. This report focuses on this emerging threat. Pulling together perspectives and data from law enforcement, tech companies, civil society, and caregivers, this report aims to provide a comprehensive overview of the escalating danger and suggest potential mitigation strategies.

The insights in this report were derived through qualitative and quantitative research methodologies:

- Aggregation of existing data and publications: review and collection of existing published data and literature.
- News reporting: thorough review of news reports on the topic of AI-generated CSAM.
- Expert interviews: interviews with 17 experts from the public and private sector, law enforcement, and civil society.

FIGURE 1⁶

One out of eight children around the world have been victims of online child sexual exploitation.



⁶ Andy Gregory. "Prioritise children's online safety at election to tackle 'hidden pandemic' of sexual abuse, experts urge." *The Independent*. June 2024.

- Internal investigations and analysis: analysis and review of legislation and policy, and open access data sources, as well as original data collection and analysis of surveys of law enforcement and caregivers.

Two anonymous surveys were conducted by Value for Good in spring 2024 for inclusion in this report. The survey of global law enforcement was conducted in April and May of 2024 and includes the perceptions and experiences of 107 law enforcement officers in 28 countries. The survey of caregivers (e.g., parents, teachers, educational professionals, healthcare professionals) was conducted in May and June of 2024 and includes the data of 103 respondents from 15 countries. While the survey results make no claim of global representation, they do give an indication of the major challenges facing these two stakeholder groups in addressing AI-generated CSAM.

Terminology Used in this Report

In line with major actors working in the child safety field, such as the Internet Watch Foundation (IWF), the Luxembourg Guidelines, and the National Center for Missing and Exploited Children (NCMEC), this report endeavors to use the most up-to-date terminology, acknowledging limitations where present.⁷

Child Sexual Abuse Material (CSAM) is any representation, by whatever means, of a child engaged in real or simulated explicit sexual activities or any representation of the sexual parts of a child for primarily sexual purposes.⁸

- **CSAM** is used to refer to CSAM that has not been manipulated by AI. As there is no widely accepted term to refer to CSAM not manipulated by AI, this choice has been made to reflect the severity of both AI-generated and non-AI-generated CSAM. Terms will be clarified as necessary throughout the report.

- **AI-generated CSAM** is all CSAM that has been manipulated or created using generative AI. This can include anything from certain images of children created by “nudify” apps to AI-generated images and videos that contain children in sexually explicit situations. In some definitions of AI-generated CSAM, including that of the National Center for Missing and Exploited Children (NCMEC), this includes text that is meant to simulate sexually explicit conversations with children or to generate manuals for child sexual exploitation.⁹

Contact offending, or contact-driven offending, refers to sexual perpetrators who seek in-person contact with their child victims to engage in physical sexual abuse.¹⁰

Deepfakes is a term that blends the terms “deep learning” and fake. Deepfakes are a subset of generative AI’s capabilities. Deepfake typically refers to media files (including images, videos, and audio) that have been crafted using AI to convincingly replace one individual’s likeness with someone else’s, often with the intent to deceive.

Foundation model refers to the largest general-purpose models that can support a diverse range of use cases. They are trained with a large set of data and can be used for different tasks, with limited fine-tuning.

⁷ The Luxembourg Guidelines are an initiative by international partners to harmonize terms related to child protection. ECPAT International, “Luxembourg Guidelines: Terminology Guidelines for the Protection of Children from Sexual Exploitation and Sexual Abuse,” Interagency Working Group on Sexual Exploitation of Children, January 28, 2016.

⁸ UNICEF, “Ending Online Sexual Exploitation and Abuse: Ensuring Children’s Digital Safety,” New York: UNICEF, 2021.

⁹ Shehan, “Addressing Real Harm.”

¹⁰ Peter Briggs, Walter T. Simon, & Stacy Simonsen, “An Exploratory Study of Internet-Initiated Sexual Offenses and the Chat Room Sex Offender: Has the Internet Enabled a New Typology of Sex Offender?,” *Sexual Abuse: A Journal of Research and Treatment*, No. 1 (2011):23, p. 72-91.

Layers of the web:

- **Clear or surface web** is the region of the internet that is easily navigable on search engines and contains publicly available web pages.
- **Deep web** is the region of the internet that is unavailable for public access because it is not indexed by search engines, but not typically used for malicious activities. This includes, but is not limited to: email, subscription content, and internal company networks.
- **Dark web** includes hidden criminal websites and services hosted on darknet networks to intentionally obscure access and enable illicit activities. This includes the buying and selling of illegal items, such as drugs, weapons, and pornography, on dark websites like Silk Road, which was widely acknowledged as the first modern dark web marketplace and eventually shut down by the FBI.

“Nudify” apps allow users to “undress” people depicted in photographs or videos using generative AI, thereby creating non-consensual nude images. These apps almost exclusively work on women and are often used by children on photos of their female classmates.

Online child sexual exploitation and abuse includes an evolving range of practices: CSAM, grooming children online for sexual purposes, live streaming of child sexual abuse and other related behaviors such as sexual extortion, the non-consensual sharing of self-generated sexual content involving children, and unwanted exposure to sexualized content.¹¹ The term is also used to refer to all types of child sexual exploitation and abuse that is partly or entirely facilitated by technology (internet or other wireless communications).¹²

Perpetrator: This report uses the term “perpetrator” instead of “offender” so as not to assign criminality to individuals who have not been found guilty in a court of law. In the event of a court decision classifying someone as an offender, this report adopts that terminology.

Safety by design is an approach that places safety at the center of the design and development process of digital tools and experiences. Its goal is to minimize online threats by anticipating, detecting, and eliminating harm before it occurs. This often necessitates changes to product design, company culture, and profit structures to prioritize child safety.¹³

Self-generated material, or self-generated sexual content/material involving children, refers to CSAM that is created by a child of themselves, under coercion or voluntarily. Of the voluntarily created CSAM, these images are often shared by former partners for blackmail or humiliation.

Sexual extortion, or sextortion, refers to a form of extortion in which individuals are blackmailed using either images they have shared or that have been created by AI to extort sexual favors or for financial gain, under threat of sharing the images on social media or with family members.¹⁴ This report focuses on the impact of sextortion on children. This definition builds on the understanding of sextortion outlined in the 2022 “Gaming and the Metaverse” report.¹⁵

Sharenting, a portmanteau of “sharing” and “parenting,” describes the trend of parents publicizing a significant amount of potentially sensitive content—images and videos—of their children on the internet and especially on social media.

¹¹ ECPAT International, “Intervention on Cybercrimes Against Children, Including CSAM,” United Nations Office on Drugs and Crime, N.d.

¹² UNICEF, “Ending Online Sexual Exploitation.”

¹³ Bracket Foundation, “Gaming and the Metaverse: The Alarming Rise of Online Sexual Exploitation and Abuse of Children Within the New Digital Frontier,” 2022.

¹⁴ Bracket Foundation, “Gaming and the Metaverse.”

¹⁵ Bracket Foundation, “Gaming and the Metaverse.”

I. “We Cannot Arrest Our Way Out of this Problem”¹⁶

THE EMERGING THREAT OF AI-GENERATED CSAM

What Is Generative AI?

AI systems are computer systems that use algorithms to replicate human abilities with a certain degree of autonomy. While there are a variety of AI systems, the most well-known are based on machine learning algorithms, which learn from examples and not from specific human instructions.¹⁷ While AI systems have been in development for decades, generative AI truly moved into mainstream consciousness in 2022 with the release of OpenAI’s ChatGPT, a chatbot and virtual assistant based on a large language model. What sets generative AI apart from earlier examples of artificial intelligence is that it can create new content.¹⁸ Generative AI is underpinned by deep learning, a powerful subset of machine learning that is based on neural networks.¹⁹

ChatGPT and similar generative models are sophisticated systems with a simple goal: to predict the next word in a sentence. For that prediction they analyze the dependencies and combinations of the preceding words, to supply the most likely fit.²⁰ Generative AI models can make these suggestions as they have been trained on large quantities of texts, learning the most common patterns, combinations and sequences of words. Based on that training, after receiving a prompt, they can create original content by outputting the most probable sequence of words. These are known as large language models, which are defined by their ability to create, manipulate, and understand text in a way that is human-like.²¹

Importantly, once a generative AI model has been trained on a dataset, it cannot learn anything that is outside the scope of its training data.

This means that an individual model does not currently possess unfettered generative abilities and is inherently limited by the training dataset. To learn new things, it needs to be re-trained on new and updated data. However, many companies continue to work towards building artificial general intelligence systems, which are not trained for a specific purpose and have the ability to learn and adapt to new tasks.²²

Generative AI can produce more than just text. These models can create images and even videos through the sequencing of random pixels. A specific type of these AI models—diffusion models—begins with an image of random pixels and in each iteration, it outputs less random matrices of pixels, creating more defined shapes.

Diffusion models require particularly large-scale datasets, usually made up of images that are scraped from the internet and then tagged so that the model can learn what is depicted in each image. This training data comes from a variety of sources as a web crawler finds URLs on the web and a web scraper extracts the data. Web crawling and scraping is one method to create large-scale datasets, as the order of magnitude of images required to train a diffusion model is in the hundreds of millions. Ownership of these datasets typically lies with the entity that curates and compiles them. The rise of AI has led to a growing dataset market in which companies sell or license access of their datasets to third parties.²³ Platforms like Google Dataset Search, Kaggle, and AWS Data Exchange provide marketplaces where datasets can be bought, sold, or accessed under various licensing agreements.²⁴

¹⁶ Phil Attwood & Simon Bailey, interview by Value for Good, May 16, 2024.

¹⁷ UN Interregional Crime and Justice Research Institute (UNICRI), “Introduction: Responsible AI Innovation,” 2024.

¹⁸ Adam Zewe, “Explained: Generative AI,” *MIT News*, November 9, 2023.

¹⁹ UNICRI, “Introduction: Responsible AI Innovation.”

²⁰ Zewe, “Explained: Generative AI.”

²¹ Deepthi Sudharsan, “Decoding the Relationship: Language Models and Natural Language Processing,” *Medium*, August 20, 2023.

²² UNICRI, “Introduction: Responsible AI Innovation.”

²³ Grand View Research, “AI Training Dataset Market Size, Share and Trends Analysis Report, By Type (Text, Image/Video, Audio), By Vertical (IT, Automotive, Government, Healthcare, BFSI), By Regions, And Segment Forecasts, 2023–2030,” accessed June 10, 2024.

²⁴ Grand View Research, “AI Training Dataset.”

How Is Generative AI Used to Create CSAM?

There are two ways for generative AI to create CSAM. To be able to create CSAM, generative AI models must either have been trained on adult pornography and extrapolate CSAM from this or have been trained on CSAM. In both cases, the content in the training datasets is critical. Training datasets can be created to meet specific parameters—with developers adding or deleting specific information—but are also created through the unrestricted scraping of images from the web.

The first widely reported instance of generative AI being used to create CSAM came in December 2023, when the Stanford Internet Observatory determined that Stable Diffusion 1.5—a latent text-to-image diffusion model originally released by Runway ML and later purchased by Stability AI—was able to generate explicit material of children because it had been trained on the LAION-5B dataset which included CSAM.²⁵ Prior to this discovery, experts in the field assumed that AI-generated CSAM was only extrapolated from the inclusion of adult pornography in training datasets.²⁶

The LAION-5B dataset was found to have been “fed by essentially unguided crawling.”²⁷ This resulted in the inclusion of a substantial volume of explicit content, including CSAM.²⁸ The exact reason why CSAM was picked up in the crawl is unknown. The LAION-5B dataset is a large-scale open-source dataset consisting of 5 billion image-text pairs scraped from the web.²⁹ LAION is a German non-profit organization that retrieves its dataset from CommonCrawl, an openly accessible repository of web-scraped data. LAION-5B is used extensively in training AI models for tasks such as image generation and language understanding. Though Stable Diffusion is the most high-profile example, other text-to-image models, including Midjourney from Midjourney, Inc., have also been trained on LAION-5B.³⁰

A Human Rights Watch report was able to trace images used by LAION-5B to specific children in Brazil, showing that many of the photos used as inputs were taken from personal family blogs.³¹ While not CSAM, the images of these children, when combined with adult pornography or other CSAM included in the datasets, could train a model on how to create AI-generated CSAM, calling into question the broader practices that are used to create these training datasets.

It was clear to perpetrators from the outset that these technologies had the potential to generate CSAM.³² In chatrooms, law enforcement observed that perpetrators were considering early on how to take models offline to avoid detection.³³

The Stable Diffusion model was released publicly without any restrictions on what to generate, likely due to Stability AI’s stated commitment to open-source AI research and the democratization of machine learning.³⁴ Without restrictions and working in an offline environment, perpetrators’ ability to create hyper-realistic AI-generated CSAM increased.

Since the news broke that Stability AI’s training model included CSAM, there have been widespread calls for developers to remove illegal and potentially harmful content from their training datasets. Promisingly, Stability AI’s subsequent models have not included CSAM in their datasets and were restricted to not be able to generate CSAM. However, models without these restrictions can still be found and circulated by perpetrators. This example of Stability AI in 2023 is indicative of a larger problem of a lack of safety-by-design in tech releases, which will affect children for years to come.

25 David Thiel, “Identifying and Eliminating CSAM in Generative ML Training Data and Models,” Stanford Internet Observatory, 2023; Alexandra Levine, “Stable Diffusion 1.5 Was Trained on Illegal Child Sexual Abuse Material, Stanford Study Says,” *Forbes*, March 25, 2024.

26 Thiel, “Identifying and Eliminating CSAM.”

27 Thiel, “Identifying and Eliminating CSAM.”

28 Thiel, “Identifying and Eliminating CSAM.”

29 Romain Beaumont, “LAION-5B: A New Era of Open Large-Scale Multimodal Datasets,” LAION.ai, March 31, 2022.

30 Levine, “Stable Diffusion 1.5 Was Trained.”

31 Human Rights Watch, “Brazil: Children’s Personal Photos Misused to Power AI Tools,” HRW, June 10, 2024.

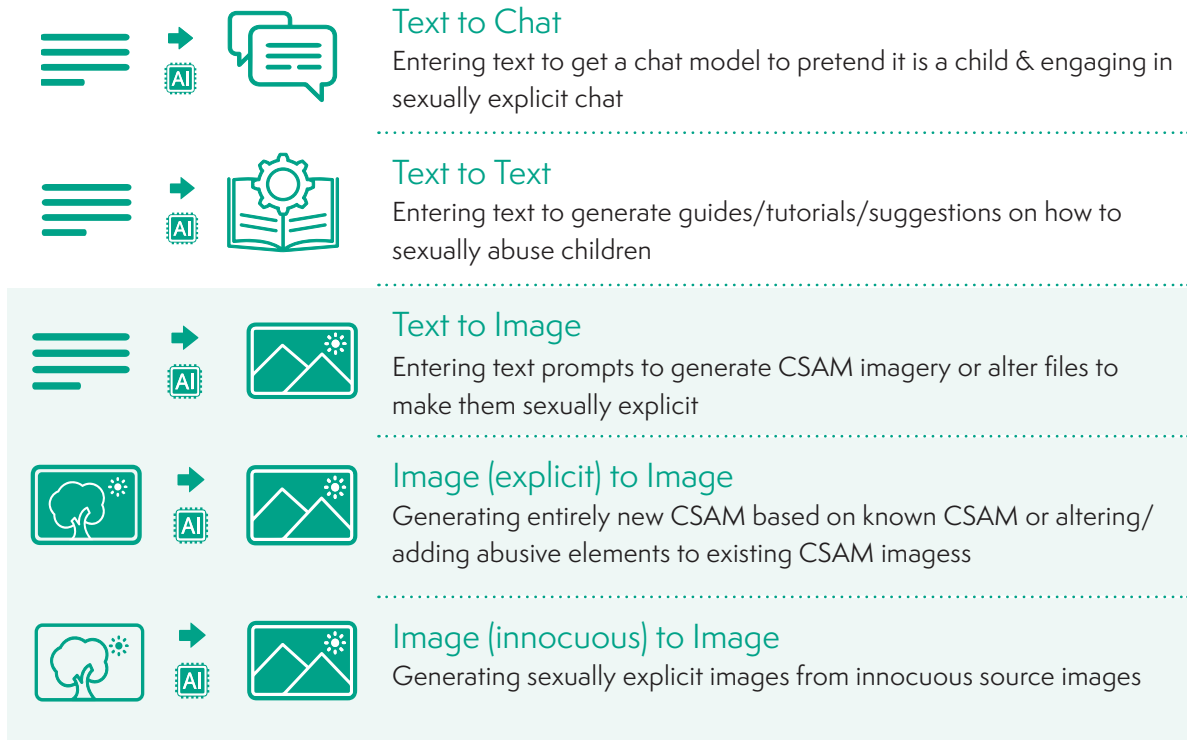
32 Mike Frend, interview by Value for Good, April 24, 2024.

33 Frend, interview.

34 Stability AI, “Stable Diffusion Launch Announcement,” stability.ai, August 10, 2023.

FIGURE 2: OVERVIEW OF TYPES OF AI-GENERATED CSAM³⁵

AI-generated CSAM isn't just still imagery, but a range of content...



Increasingly including moving imagery

What Types of CSAM Can Generative AI Create?

Generative AI has enabled the creation of a wide range of CSAM content. The type of AI-generated CSAM that has received the most attention is “text-to-image,” meaning that a perpetrator gives an AI model a prompt, directing the AI model to generate or alter an image to be sexually explicit in its depiction of a child. The photorealistic nature of text-to-image AI-generated CSAM has matured quickly, with perpetrators now able to create images that in many cases are indistinguishable from CSAM.

Text-to-chat and text-to-text AI-generated CSAM are also very advanced. With text-to-chat AI-generated CSAM, users can find themselves in increasingly escalating situations due to a lack of safeguards. Users may also intentionally simulate sexually explicit or exploitative conversations with children. Some of these chatbots, colloquially known as “AI girlfriends,” are already designed to engage in sexually explicit conversations and simulate romantic relationships with users. Because many of these “AI girlfriends” do not have strong safeguards, some have been shown to lead users towards increasingly escalating content and behavior, including seeking out CSAM.³⁶ Perpetrators can also intentionally manipulate such AI chatbots to interact as

³⁵ Shehan, “Addressing Real Harm Done by Deepfakes”, video, accessed April 20, 2024.

³⁶ Stability AI, “Stable Diffusion Launch Announcement”, stability.ai, August 10, 2023.

a child might when engaging in sexually explicit or exploitative conversations. A particular danger of these chatbots, notes U.K. Online Child Sexual Exploitation and Abuse Covert Intelligence Team (OCCIT) supervisor Mike Frend in an interview for this report, are the lack of in-built safeguards:

“Chatbots are very proactive in leading users to this explicit content; chatbots will give suggestions on how to commit crimes, encourage users to harm themselves, and keep drawing users in. When we speak to offenders post-conviction, they often remark on their lack of resilience to turn away from the temptation to offend, especially when the chatbots are so enticing.”

Text-to-text generative AI models are often unable to filter out what they should not produce, beyond a list of words that they should not generate.³⁷ For instance, when a popular generative AI model was asked for a list of websites where pirated movies are available for download, the model replied that it could not provide such a list. However, when the reverse of the question was asked—namely, a request for a list of websites to avoid if one wants to make sure they are not pirating movies—the model returned a list of websites.³⁸ Some models that lack further safeguards are even able to produce “how-to” guides on extorting victims using CSAM or AI-generated CSAM.³⁹

Image-to-image AI-generated CSAM is also growing in popularity and in capability. These developments are particularly concerning, with perpetrators able to feed two to three still images of a child into an AI model and refine it within an hour and then create tailored AI-generated CSAM of that child within seconds.⁴⁰ This capability enables the exploitation of innocuous images of children, creating victims, who may never know their images were used for CSAM. Moreover, image-to-image CSAM can revictimize existing victims by “improving” the quality of known CSAM or altering images to have the existing victim engage in additional sexual acts, specific to a perpetrator’s fantasies.

Recent developments in AI have also allowed for the creation of photorealistic moving imagery, which is rapidly becoming indistinguishable from CSA videos. For instance, while in late-2023, law enforcement could identify AI-generated CSA videos by the absence of blinking eyes, perpetrators have been able to eliminate this tell. The Internet Watch Foundation (IWF) also predicted in July 2024 that future AI-generated CSA videos will be of a “higher quality and realism.”⁴¹

Law enforcement has noted that the rapid development of AI-generated CSAM is fueled by the “supportive” nature of the perpetrator community.⁴² One officer interviewed for this study remarked that perpetrators “support others to help improve prompts and make ‘better quality’ AI-generated CSAM.”⁴³ They exchange advice on avoiding detection, handling police inquiries, and using AI-generated images for blackmail or “sextortion.”⁴⁴ This collaboration allows them to refine tactics, stay ahead of law enforcement, and evade capture.

Particularly troubling is how easy it is to find these communities on such mainstream social media platforms as TikTok and Instagram.⁴⁵ Perpetrators who are producing, viewing, and commercializing AI-generated content, often tease at the high-quality of their material over mainstream platforms like Instagram and direct interested users to interact with them over other channels—including Telegram or dark web sites—where no safeguards exist to prevent the trade of material.⁴⁶

Journalists have reported seeing users asking on Instagram and TikTok which AI models were used to produce various CSAM images and posting instructional videos on TikTok that show followers how to “generate and perfect” explicit photos of young girls.⁴⁷ Despite the fact that it is still in its nascency, generative AI has quickly evolved to facilitate nefarious activity and has increased mainstream access to CSAM.

37 Tom Oldroyd, interview by Value for Good, June 13, 2024.

38 Oldroyd, interview.

39 Frend, interview.

40 Frend, interview.

41 Dan Milmo, “AI Advances Could Lead to More Child Sexual Abuse Videos, Watchdog Warns,” *The Guardian*, July 22, 2024.

42 Frend, interview.

43 Frend, interview.

44 Alex Hern, “Can AI Image Generators be Policed to Prevent Explicit Deepfakes of Children?,” *The Guardian*, April 23, 2024; Attwood & Bailey, interview.

45 Frend, interview; Alexandra Levine, “I Want That Sweet Baby: AI-Generated Kids Draw Predators On TikTok And Instagram,” *Forbes*, May 20, 2024.

46 Frend, interview.

47 Levine, “I Want That Sweet Baby.”

Estimating the Prevalence of CSAM and AI-Generated CSAM

Due to the concealed nature of the creation and trade of CSAM, it is hard to quantify the amount of material that is created, possessed, and shared. Still, all indicators point to a growing volume of content, indicating that more children are being harmed through the creation of CSAM.

Estimates currently point to forced commercial sexual exploitation being a \$172-billion industry, roughly the size of Kuwait's 2024 gross domestic product.⁴⁸ The effects of sexual exploitation are widespread, with one out of eight children around the world having been victims of online child sexual exploitation (i.e. non-consensual taking, sharing and exposure to sexual images) in 2023.⁴⁹ A U.K. study found that up to 14 million U.S. men (11% of the male population) have engaged in online sexual abuse of children.⁵⁰ Additionally, over 850,000 men in the UK reportedly have a sexual interest in children, highlighting the urgent need to protect children from exploitation.⁵¹

International data from the National Center for Missing & Exploited Children (NCMEC) show a significant increase in the amount of CSAM available online and reported via online platforms.⁵² Analyzing available data from 2010–2023, the number of CyberTipline reports related to child sexual abuse has skyrocketed, reaching 36 million reports of suspected child sexual exploitation in 2023.⁵³ IWF reported in 2021 the largest increase in new CSAM of children between the ages of eleven and thirteen years old, namely a 75% increase

from 2020 in volume of material.⁵⁴ While not an exact count of the prevalence of CSAM, all available data point to an increasingly dangerous environment for children.

Law enforcement report seeing a similar explosion in the volume of CSAM collected during raids.⁵⁵ They attribute this rise to a variety of developments, including increased ease of access, increasingly secure ways to store CSAM digitally, and cheap storage options. Historically, CSAM has been primarily hosted on the dark web, creating a barrier to access. However, the increasing distribution of CSAM through peer-to-peer networks and the deep web has made it more accessible to less technologically savvy perpetrators.⁵⁶



48 International Labour Organization, "Profits and Poverty: The Economics of Forced Labour," Second edition, March 19, 2024; Worldometer, "GDP by Country," accessed May 20, 2024.

49 Andy Gregory, "Prioritise Children's Online Safety at Election to Tackle 'Hidden Pandemic' of Sexual Abuse, Experts Urge," *The Independent*, June 2, 2024.

50 "Into the Light," Childlight Global Child Safety Institute; Office for National Statistics, "Population Estimates for the UK, England, Wales, Scotland, and Northern Ireland: Mid-2022," March 26, 2024.

51 National Crime Agency, "NCA Leads International Coalition Tackling Child Sexual Abuse," 2023.

52 NCMEC is a U.S.-based non-profit organization that manages the CyberTipline, a database of child sexual exploitation, and supports victims and law enforcement in removing explicit content of children from the Internet

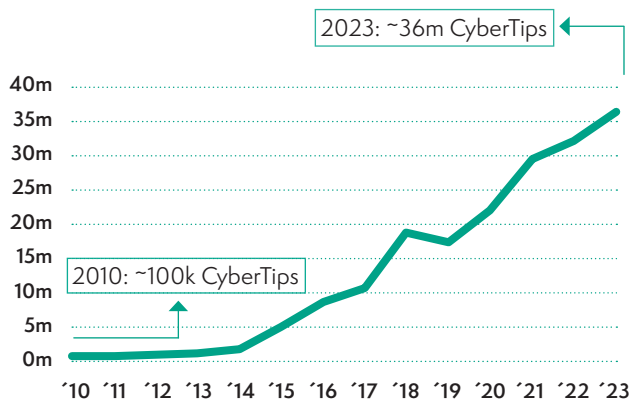
53 NCMEC has been running their CyberTipline since 1998; National Center for Missing & Exploited Children; NCMEC, "CyberTipline 2023 Report." Note: Includes all CSAM, real & AI-generated. Data only accessible from 2010. Between 2019, when "Artificial Intelligence: Combating Online Sexual Abuse of Children," the first Bracket Foundation study was published, and today, there has been a 210% increase in the reported amount of CSAM.

54 Internet Watch Foundation, "Total Number of Reports," 2021.

55 Internet Watch Foundation, "Total Number of Reports," 2021.

56 Attwood & Bailey, interview; David Haddad, interview by Value for Good, May 29, 2024.

FIGURE 3: RISE IN NCMEC CYBERTIPLINE REPORTS⁵⁷



Number of NCMEC CyberTipline Reports Related to Child Sexual Abuse (2010 – 2023)

This increase in content is also attributed to the ease of access to affordable and abundant digital storage. Though some perpetrators may purge material periodically out of guilt or to avoid detection, many are now digitally siloing their collections, giving them a sense of security and likely encouraging longer retention.⁵⁸

The rise of CSAM is also at least partially due to young people's behavior online. Youth are increasingly active online and their willingness to share explicit images of themselves on the web or through encrypted digital platforms has exacerbated the situation. This "self-generated" material can be created under coercion, in at-the-time consensual relationships, or out of self-interest. Law enforcement has seen a shift from CSAM created through abuse a decade ago to a significant portion now being "self-generated," whether through coercion or consensually.⁵⁹ Among CSAM analyzed by IWF in 2021, 49% of CSAM depicting seven- to ten-year-olds was "self-generated," while the figure among sixteen- to seventeen-year-olds was 92%.⁶⁰ This multitude of factors helps to explain the dramatic increase in CSAM.

The prevalence of AI-generated CSAM is even more challenging to estimate. In 2023, NCMEC began collecting data on the number of CyberTipline reports that contained verified AI-generated CSAM, coming from both AI platforms and social media platforms.

Understanding NCMEC's CyberTipline⁶¹

NCMEC's CyberTipline is one of the most powerful tools available to report online child sexual exploitation and abuse, identify perpetrators, and grasp the magnitude of the issue. Still, the organization is not always able to act as an effective clearing house. A recent report from the Stanford Internet Observatory, identified key weaknesses in NCMEC's operations, which include the possibility that its API incentivizes platforms to overreport. Two challenges may impact the reliability of the CyberTipline's data, though do not undermine the conclusion that online child sexual exploitation and abuse is at an all-time high:

1. Platforms err on the side of caution in reporting images (e.g., reporting an explicit image of a 22-year-old who looks 17-years-old), which overloads the system with extraneous reports.
2. Meme—or viral—content overwhelms platforms and the CyberTipline, as viral content is reported repeatedly, often with no repercussions for the posters.

⁵⁷ National Center for Missing & Exploited Children, CyberTipline 2023 Report, 2023.

⁵⁸ Attwood & Bailey, interview;

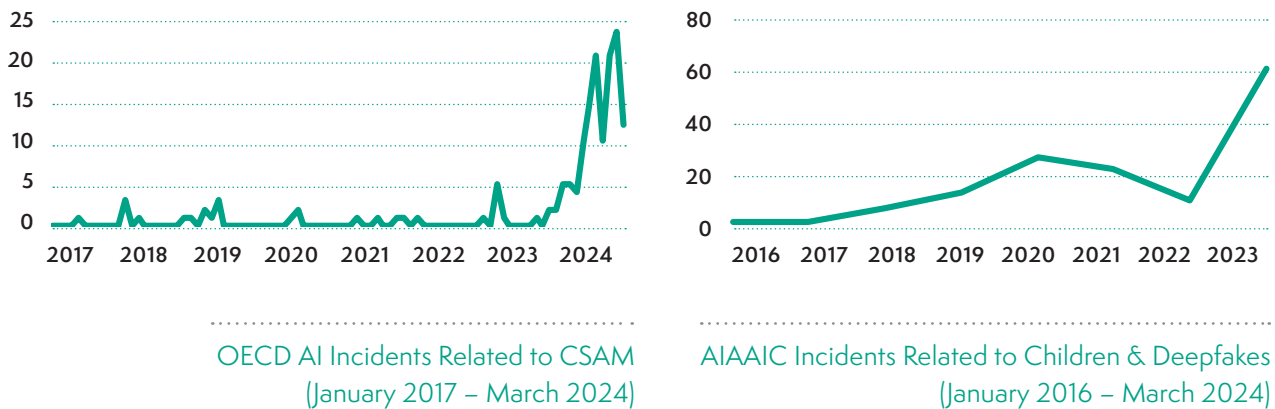
⁵⁹ Attwood & Bailey, interview.

⁶⁰ Internet Watch Foundation, "Self-Generated Child Sexual Abuse," 2021.

⁶¹ Shelby Grossman et al., "The Strengths and Weaknesses of the Online Child Safety Ecosystem," Stanford Internet Observatory, 2024.

FIGURE 4: RISE IN AI INCIDENTS RELATED TO CSAM⁶²

Two databases that track AI incidents show increased incidence of AI-generated CSAM & related deepfakes



As of spring 2023, just five generative AI platforms have registered to report to CyberTipline and submit reports, despite all U.S.-based Electronic Service Providers (ESPs) being legally required to report instances of CSAM to the CyberTipline once they become aware of them. The low number of registrations could be due to the lack of visibility that these platforms have regarding the use of their models to create AI-generated CSAM. As discussed above, once a model is taken offline, the platform has no oversight as to what is created.

Of the approximately 36 million CyberTipline reports in 2023, 4,700 included verified AI-generated CSAM.⁶³ Traditional online platforms, like Meta and X, were where the majority of verified AI-generated CSAM was found in 2023, with more than 70% of AI-generated CSAM tips submitted through these platforms.⁶⁴

In addition to NCMEC, other databases have been tracking the increased incidence of AI-generated CSAM and deepfakes. The OECD AI Incidents and the AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC) databases both show a stark uptick in AI-related child sexual abuse, child grooming, deepfakes, and deepnudes in 2023, compared to the handful of incidents each year between 2016 and 2022. Other measures of similar content also saw a peak in 2023; Home Security Heroes, an identity theft start-up, found that in 2023 the total number of deepfake videos online had reached 95,820, up 550% from 2019. 98% of these deepfake videos were pornography and 99% of the individuals targeted in deepfake pornography videos were women.⁶⁵

⁶² Left Graph: OECD, "OECD AI Incidents Monitor," accessed April 11, 2024. Right Graph: AIAAIC, "AIAAIC Repository," accessed April 11, 2024.

⁶³ Shehan, "Addressing Real Harm." The list of AI platforms includes: BashTech LLC (402 CyberTips), OpenAI (329 CyberTips), Gab AI Inc. (2 CyberTips) and Anthropic (1 CyberTip).

⁶⁴ Shehan, "Addressing Real Harm."

⁶⁵ Home Security Heroes, "2023 State of Deepfakes. Realities, Threats, and Impact," 2023.

In the United States currently, law enforcement finds AI-generated materials in approximately 50% of their seizures.⁶⁶ As of spring 2024, over 50% of law enforcement officers surveyed for this study report having encountered AI-generated CSAM in their work, including “nudify” apps used on children, AI-generated CSAM depicting famous individuals, and AI used to alter CSAM of existing victims.⁶⁷

Chainalysis, a blockchain analysis firm, tracks the flows of cryptocurrency on the dark web. They have noted that the CSAM market appears to have been “flooded with content” in the last few years, citing that the increase in content has caused the price of CSAM to collapse. This deluge of content could potentially be driven by the presence of AI-generated CSAM.⁶⁸

While these indications of the quantity of AI-generated CSAM are helpful to understand the extent of the current problem, determining which content is AI-generated will only become more challenging as technology continues to improve.

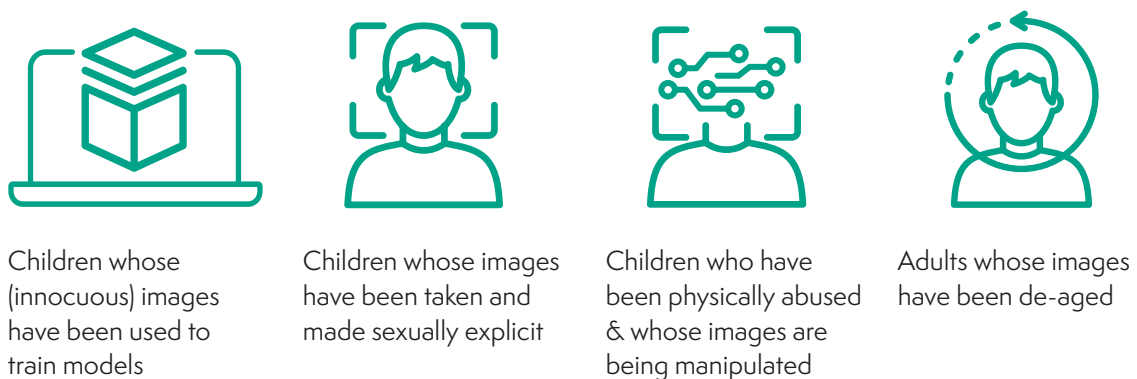
How Is AI-Generated CSAM Impacting the Landscape of Threats to Children Online?

Impact on Child Sexual Abuse Victims

Generative AI has created new ways in which online exploitation and abuse can occur and has exacerbated existing harm. As all indicators show, the total amount of CSAM is rising and the presence of AI-generated CSAM is complicating the landscape. Generative AI has introduced four new ways in which children can become victims of sexual abuse and exploitation:

- 1. Images used to train models:** Innocuous images and existing CSAM are used in the training datasets for AI models. This results in the likeness of many children being used to enable further creation of AI-generated CSAM. Stability AI’s Stable Diffusion model enabled this type of victimization, as it was trained on the LAION-5B training dataset, which included over 3,000 entries of suspected CSAM and many more innocuous images of children.⁶⁹

FIGURE 5: AI-GENERATED CSAM VICTIM ARCHETYPES⁷⁰



⁶⁶ Haddad, interview.

⁶⁷ Value for Good, survey with global law enforcement, spring 2024.

⁶⁸ Chainalysis, “The 2024 Crypto Crime Report,” 2024.

⁶⁹ Thiel, “Identifying and Eliminating CSAM.”

⁷⁰ Value for Good analysis.



2. **Innocuous images taken and made into AI-generated CSAM:** The capabilities of AI allow for any innocuous image of a child to become CSAM. Previously, parents and other caregivers had to worry about naked images of their children being made public; with generative AI, any image of a child can potentially be made into CSAM. Moreover, recent cases have shown that perpetrators can re-train generative AI models to generate CSAM from as few as two to three innocuous images of a child.
3. **Modification of existing CSAM:** A child may no longer be in an active situation of abuse and law enforcement may have attempted to seize and scrub all content of them; however, AI enables perpetrators to improve old, poorer quality content and create new content. *The Guardian* reported in June 2024 that child safety groups who track perpetrator behavior have observed a fixation among predators in creating more content of their favorite “star” victims.⁷¹ This infatuation with particular children is not a new phenomenon; Meta reported in 2020 that six videos comprised half of all CSAM “shared and re-shared” on its platforms.⁷² This poses new challenges for law enforcement finding and destroying this CSAM.
4. **De-aged images of adults:** Generative AI has also made victims of adults by “de-aging” their images. This material can be used to hurt public figures as well as private individuals, with predators creating the material for sexual gratification, blackmail, humiliation, or all three.

Another fear among many who work against this threat is that the presence of AI-generated CSAM runs the risk of distracting law enforcement from

identifiable victims, thereby diverting attention and resources from children who need safeguarding.⁷³

AI may also be used to obscure existing CSAM, with perpetrators exploiting the tells of AI-generated material and manipulating CSAM to look like it was created with AI.⁷⁴ By adding a sixth finger or stopping eyes from blinking, perpetrators hope they can create plausible deniability saying they were unaware the content was real.⁷⁵ While in many jurisdictions possession of AI-generated CSAM may still be considered illegal, by obscuring the CSAM, perpetrators may succeed in slowing down an investigation or hiding a child in need of help.

The motives behind creating AI-generated CSAM vary, with several theories emerging.⁷⁶ Some perpetrators may seek to showcase their skills by creating realistic images or fantasy environments.⁷⁷ Another hypothesis is that perpetrators use AI to make victims look “happier,” as they want to believe that there is a true relationship between themselves and the victim.⁷⁸ It could also be that creators see tremendous commercial potential in AI-generated CSAM, a hypothesis which is addressed in further detail later in this report. Additional theories, including whether perpetrators believe AI-generated CSAM is more morally palatable than CSAM or that perpetrators seek new “thrills”, remain to be tested.

It is likely that perpetrators will more easily be able to victimize more children—and adults—from the comfort of their computers. This increase in sexualization of children may reduce barriers for other interested perpetrators, kicking off a new epidemic of child sexual exploitation.

71 Katie McQue, “Child Predators Are Using AI to Create Sexual Images of their Favorite ‘Stars’: ‘My Body Will Never Be Mine Again,’” *The Guardian*, June 12, 2024.

72 McQue, “Child Predators Are Using AI.”

73 Will Oremus, “AI Is About to Make the Online Child Sex Abuse Problem Much Worse,” *The Boston Globe*, April 22, 2024.

74 Attwood & Bailey, interview.

75 Attwood & Bailey, interview.

76 Offender motives and psychology are not the primary remit of this study; however, it is important to understand primary motives to contextualize the broader issues.

77 National Law Enforcement Agency, interview by Value for Good, June 4, 2024.

78 National Law Enforcement Agency, interview.

“Whether a victim knows or doesn’t know, they are a victim”

– Mike Freund, OCCIT supervisor

Social Media's Role in Facilitating AI-Generated CSAM

The proliferation of AI-generated CSAM has resulted in CSAM no longer being confined to the dark web, but also now available in the mainstream. Perpetrators seem to consider the production and distribution of AI-generated CSAM less risky and embrace this new technology.⁷⁹ As social media facilitates the spread of AI-generated CSAM, there seems to be a normalization in viewing and creating this content, as U.K. OCCIT Supervisor Mike Frend explains:

"It is relatively easy to find these communities on mainstream social media platforms, with accounts hinting at how good their work is, modelers offering contact over Telegram or other encrypted apps to see more of their work, and the algorithm recommending children's accounts. Once a potential perpetrator finds an experienced modeler, they land very quickly on a payment site and since they were directed through social media, it doesn't feel as nefarious. In this sense, social media acts as a legitimization element."⁸⁰

Initial court cases show this pipeline of social media to encrypted messaging apps. For instance, U.S. federal court documents for the 2023/2024 case of a Wisconsin man show that he posted a "realistic AI-generated image of minors wearing bondage-themed leather clothes" on Instagram and encouraged his followers to "come check out what [they] are missing" by messaging him on Telegram.⁸¹

Journalists and law enforcement agents have noted how the algorithms that are the basis for content recommendations on mainstream social media platforms play a role in connecting perpetrators and leading perpetrators to similar content. For instance, as *Forbes* reporters were researching an article on this topic, TikTok began recommending such additional prompts to them as "ai generated [sic] boys" and "ai [sic] fantasy girls."⁸² A British investigative team also observed the Ins-

tagram algorithm suggesting new accounts—such as accounts of young girls—to older, male users who had shown interest in that content before.⁸³ The same phenomenon has been observed with violent and misinformative content on Facebook.⁸⁴

With social media's unique ability to connect people with similar interests, it is unsurprising that Instagram has been identified by *The Wall Street Journal* as "the most important platform" for the trade of CSAM.⁸⁵ Social media algorithms that recommend exploitative and potentially illegal content to predators, endanger children. With the potential for any innocuous image of a child on social media to become AI-generated CSAM and social media facilitating connections between a predator and the content he is interested in; social media companies will need to acknowledge their role in the spread of AI-generated CSAM online and adjust their practices to protect children.

Commercializing Content and Sextortion

The existence of AI-generated CSAM has led to a rise in the commercialization of child sexual exploitation. In addition to the buying and selling of content and models, new methods of extortion—or sextortion—have proliferated thanks to the ease with which perpetrators can now create CSAM. There are three primary drivers for perpetrators who possess AI-generated CSAM, which determine the commercial drivers:

- 1. Sexual gratification:** An interest in AI-generated CSAM for sexual gratification purposes leads to the buying and selling of both existing AI-generated CSAM as well as "customized" content. It has been observed by law enforcement that the audience for this content is typically men with a prior history of sexual interest in CSAM and an interest in how technology can enable their predatory behavior.⁸⁶

⁷⁹ Attwood & Bailey, interview.

⁸⁰ Frend, interview.

⁸¹ Drew Harwell & Pranshu Verma, "In Novel Case, U.S. Charges Man with Making Child Sex Abuse Images with AI," *The Washington Post*, May 21, 2024.

⁸² Levine, "I Want that Sweet Baby."

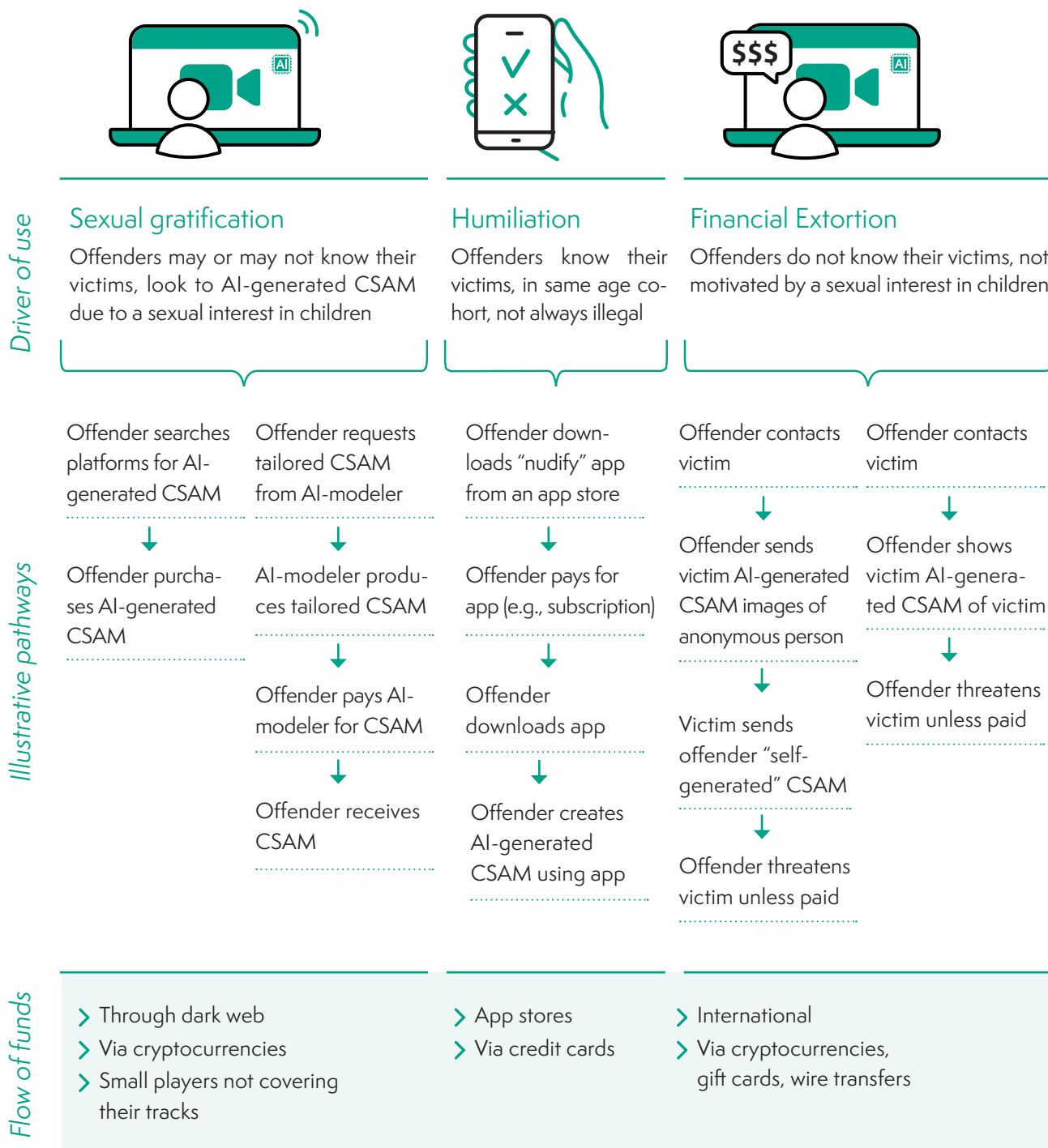
⁸³ Frend, interview.

⁸⁴ Elizabeth Dwoskin & Gerrit De Vynck, "Facebook's Internal Chat Boards Show Politics Often at Center of Decision Making," *The Washington Post*, October 24, 2021.

⁸⁵ Katherine Blunt & Jeff Horwitz, "Instagram Connects Vast Pedophile Network," *The Wall Street Journal*, June 7, 2023.

⁸⁶ Frend, interview.

FIGURE 6: DRIVERS OF COMMERCIALIZATION OF AI-GENERATED CSAM⁸⁷



⁸⁷ Value for Good analysis.



If existing AI-generated CSAM does not satisfy perpetrators, they can also request specific material or the creation of AI models from more sophisticated developers. These requests are often based on children they know. In the United Kingdom, for instance, law enforcement has identified perpetrators who create tailor-made AI-generated CSAM and specific AI models of individual children based on requests from fathers, uncles, and neighbors.⁸⁸

2. Humiliation: Perpetrators humiliate victims, from celebrities to acquaintances, by using generative AI capabilities. One common tactic is through “nudify” apps. Perpetrators will download “nudify” apps from app stores (for example, the Google Play Store or Apple’s App Store) and create nude images of their victims. Often this is done by school aged peers.⁸⁹ Although the apps claim to prohibit use of images of children, many reports indicate that they do not stop users from uploading such images (for more information, see deep dive at the end of this section). Most of these apps offer a “freemium model,” in which the first few images can be created for free and further images require a subscription to the app. As of June 2024, four “nudify” apps compared by AI Mojo cost between \$2.39 and \$14.99 per month.⁹⁰ Analysis of 34 “nudify” apps by Graphika showed that they had over 24 million visitors in September 2023.⁹¹

3. Financial extortion: Financial motivation drives further commercialization of AI-generated CSAM. Generative AI has also changed the landscape of sextortion, or the practice of extorting money or sexual favors from someone by threatening to release explicit images or information about them.

Historically, perpetrators needed to coerce victims into providing self-generated images, but with generative AI, they can create explicit images from innocuous photos and threaten to release them.⁹² Sextortion, of all forms, is a major threat, with NCMEC receiving an average of 812 reports of sextortion weekly over 2023.⁹³ NCMEC and Thorn, an international organization that works to address the sexual exploitation of children, predict that the use of generative AI may increase these cases in the coming years.⁹⁴

Catfishing, where perpetrators use a fictional online persona to coerce someone into a relationship or provide explicit imagery, has long been a tactic for sextortion. However, AI-generated CSAM has expanded these tactics. There is an increasing number of reports of scammers contacting teenagers over Instagram or other social media platforms, using AI-generated images of girls to coerce teenage boys into sending illicit photos of themselves. U.S. crime statistics show that cases of such sextortion rose to 26,700,000 in 2023, a figure that is likely underreported due to victims’ shame, fear of being blamed for their victimization, or fear that the perpetrator will carry out threats.⁹⁵

Initial reports by the *BBC*, indicate that Nigeria is a hotspot for attackers, with many Nigerian TikTok, YouTube, and Scribd accounts exchanging recommendations and templates for the extortion of victims.⁹⁶ Based on data from the FBI and NCMEC, sextortion is the “most rapidly growing crime” affecting children in the United States, Canada, and Australia; at least 27 suicides have been linked to victims of sextortion in the United States, a figure that is likely to grow as illicit content becomes even easier to produce.⁹⁷ The global nature of these crimes raises particular challenges for law enforcement, a theme that will be discussed in section II.

⁸⁸ Frend, interview.

⁸⁹ Natascha Singer, “Teen Girls Confront Epidemic of Deepfake Nudes at School,” *The New York Times*, April 8, 2024.

⁹⁰ Roopal, “AI-Generated Child Sexual Abuse: The Threat of Undress AI Tools,” AI Mojo, June 2024.

⁹¹ Santiago Latakos, “A Revealing Picture,” *Graphika*, December 2023.

⁹² Shehan, “Addressing Real Harm.”

⁹³ Thorn & NCMEC, “Trends in Financial Sextortion: An Investigation of Sextortion Reports in NCMEC CyberTipline Data,” June 2024.

⁹⁴ Thorn & NCMEC, “Trends in Financial Sextortion.”

⁹⁵ Joe Tidy, “Dead in 6 Hours: How Nigerian Sextortion Scammers Targeted my Son,” *BBC*, June 9, 2024; U.S. Department of Justice, “Sextortion, Crowdsourcing, Enticement, and Coercion,” 2023.

⁹⁶ Tidy, “Dead in 6 Hours.”

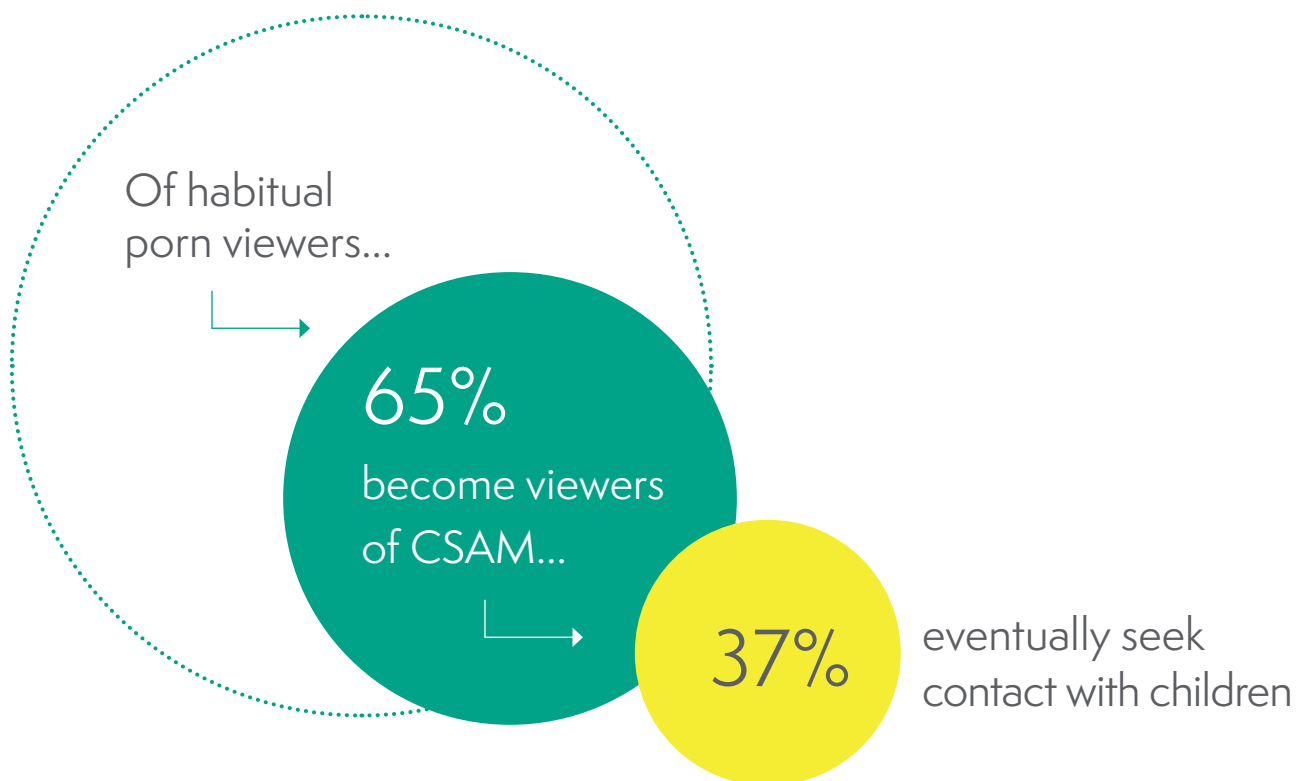
⁹⁷ Paul Raffile et al., “Digital Pandemic: Uncovering the Role of ‘Yahoo Boys’ in the Surge of Social Media-Enabled Financial Sextortion Targeting Minors,” NCRI, 2024; Tidy, “Dead in 6 Hours.”

Contact Offending

A looming concern in the age of AI-generated CSAM is the pathway from AI-generated CSAM to contact offending, the in-person contact with child victims by perpetrators to engage in physical sexual abuse. Research from the child’s rights organization, Protect Children, shows that exposure to pornography, and desensitization to hardcore pornography in particular, drives viewers to CSAM.⁹⁸ Often the algorithms on porn websites tacitly fuel this interest, directing viewers to CSAM without them actively searching it out.⁹⁹

Examples exist of perpetrators reporting that viewing CSAM “reinforced a sexual interest in children” and gave them the courage to engage in contact offending.¹⁰⁰ Other experts believe that CSAM has also led to a widespread acceptance of sexualizing children.¹⁰¹ AI-generated CSAM aids the “normalization” of child sexual abuse, by engaging perpetrators’ fantasies and bringing content into the mainstream.¹⁰² With most physical sexual abuse of children taking place in the familial and extrafamilial environment, as well as in peer-to-peer interactions, children are particularly at risk in environments where there is limited oversight.¹⁰³ Although it’s too early to confirm, AI-generated CSAM may have similar effects, especially as it becomes indistinguishable.

FIGURE 7: BY THE NUMBERS: RELATIONSHIP BETWEEN VIEWING PORNOGRAPHY AND CONTACT OFFENDING¹⁰⁴



⁹⁸ Tegan Insoll, et al., “CSAM Users in the Dark Web,” Protect Children, September 27, 2021; Attwood & Bailey, interview.

⁹⁹ Attwood & Bailey, interview.

¹⁰⁰ McQue, “Child Predators Are Using AI.”

¹⁰¹ United Nations and UN Human Rights, “Sale and Sexual Exploitation of Children,” 2023.

¹⁰² Caitlin Roper, “‘No Way of Knowing if the Child Is Fictional’: How ‘Virtual’ Child Sexual Abuse Material Harms Children,” Collective Shout, May 6, 2024.

¹⁰³ Attwood & Bailey, interview.

¹⁰⁴ Insoll, Ovaska, and Vaaranen-Volkonen, CSAM Users in the Dark Web, 2021.



Deep Dive: The Threat of “Nudify” Apps & Legal Implications

“Nudify” or “undressing” apps use AI to “undress” images, predominantly of women and girls. In some cases, if a man’s picture is uploaded, the app still generates female anatomy.¹⁰⁵ By filling in a “best-guess” of what the woman in the picture would look like without her clothing on, “nudify” apps create non-consensual nude images. While using these apps on images of children may not always meet the definition of AI-generated CSAM, they are still a serious and increasingly widespread danger.

Mainstream Usage: “Nudify” apps have become particularly popular among minors, with dozens of reports of teenage boys using the apps’ features on their female classmates. U.K. law enforcement estimates that at least one child in every school in the United Kingdom has one of these apps.¹⁰⁶ These apps are not confined to the dark web, rather they are available in the same app stores where children download Instagram, Snapchat, and TikTok, reducing the barriers to access.

Role of App Stores: Despite violating policies against creating non-consensual sexual images and CSAM, these apps are easily accessible on both the Google Play Store and Apple’s App Store. The stores often fail to fully vet apps before listing them, typically removing them only after media outcry or law enforcement intervention. “Nudify” apps purport to have a “legitimate” use, which includes creating nude images of consenting adults.¹⁰⁷ This raises the question of what a “legitimate” use of nudify apps is, as if someone wanted someone else to have a nude image of them, they would. The fact remains there is currently no way to guarantee that all images created with these tools involve only consenting adults.

Reports of “Nudify” Apps in Schools: The use of “nudify” apps by teenagers against their peers, often in schools, has complicated the legal response. A few initial reports and court cases involving “nudify” apps in schools have made international headlines:



- **Australia:** In Victoria in June 2024, news broke that the fake nude images of 50 girls at a secondary school were being shared on social media. A former male student at the school was arrested for creation and distribution of these AI-generated nude images, but, as of the time of this report, has yet to be charged.¹⁰⁸



- **United States:** In Florida in March 2024, two teenage boys, ages 13 and 14, were arrested for creating and sharing AI-generated nude images of their classmates and charged with third-degree felonies.¹⁰⁹ Under Florida-state law, it is a felony to share “any altered sexual depiction” without the individual’s consent.¹¹⁰ Notably, the school did not immediately expel the boys once the AI-generated CSAM came to the attention of the school administration.¹¹¹



- **United States:** In New Jersey in October 2023, reports came out about boys creating fake nude images of their classmates with “nudify” apps. Information on the repercussions they faced have not been made public, but families of the victims have said that the male student accused of creating the images was suspended for a couple days and lamented that the administration seemed to be doing everything it could to make the incident disappear.¹¹²

¹⁰⁵ Emilie Lavinia, “I’ve Seen Boys Request Fake Nudes of their Teachers and Mothers’: How Nudify Apps Are Violating Women and Girls in the UK,” *Glamour*, June 24, 2024.

¹⁰⁶ Frend, interview.

¹⁰⁷ Frend, interview.

¹⁰⁸ Rhiana Whitson, “Principals Say Parents Need to be Vigilant as Explicit AI Deepfakes Become More Easily Accessible to Students,” *ABC News*, June 25, 2024.

¹⁰⁹ Caroline Haskins, “Florida Middle Schoolers Arrested for Allegedly Creating Deepfake Nudes of Classmates,” *Wired*, March 8, 2024.

¹¹⁰ Florida Senate, Promotion of an Altered Sexual Depiction; Prohibited Acts; Penalties; Applicability, Title XLVI, § 836.13 [2023].

¹¹¹ Haskins, “Florida Middle Schoolers Arrested.”

¹¹² Julie Jargon, “Fake Nudes of Real Students Cause an Uproar at a New Jersey High School,” *The Wall Street Journal*, November 2, 2023; Singer, “Deepfake Nudes at School.”

• **Spain:** In summer 2023 reports of fake nude images circulating on social media came from the town of Almendralejo. The 28 victims were girls, aged between 11 and 17, whose pictures from their own social media accounts were taken by 15 local boys, run through “nudify” apps, and shared on WhatsApp and Telegram apps.¹¹³ The juvenile perpetrators were convicted of 20 counts of creating child abuse images and offenses against the victims’ moral integrity; their sentences include one year of probation and an order to attend classes on gender, equality, and responsibly using technology.¹¹⁴

Reactions to “Nudify” Apps: In general, schools, which often receive the first reports of “nudify” apps used by students on their peers, have treated the usage of “nudify” apps and the resulting AI-generated CSAM as a form of bullying rather than a criminal offense. Schools are generally behind in their policies for dealing with AI; while many have rules regarding the use of AI for assignments, they do not address using “nudify” apps in their codes of conduct.¹¹⁵ School administrations have been accused of making efforts to protect the perpetrators—often underage boys—from legal repercussions, rather than protecting the victims. This comes from a reticence to criminalize minors, which at the same time diminishes the severity of this crime and the harmful effects for the victims, often making victims feel unsafe in their schools following the event.¹¹⁶

This endeavor to protect juvenile perpetrators can be seen clearly with Germany’s May 2024 move to reduce the possession of CSAM from a felony to a misdemeanor. Members of the governing coalition justified the change by stating the concern that too many minors are being charged with felonies under the previous statute; however, this move also reduces the penalty for adult perpetrators.¹¹⁷

The majority of nudify apps are only trained to be able to “undress” women.

¹¹³ Guy Hedgecoe, “AI-Generated Naked Child Images Shock Spanish Town of Almendralejo,” *BBC*, September 24, 2023.

¹¹⁴ Sam Jones, “Spain Sentences 15 Schoolchildren Over AI-Generated Naked Images,” *The Guardian*, July 9, 2024.

¹¹⁵ Kate Linebaugh, host, “Teens Are Falling Victim to AI Fake Nudes,” *WSJ Podcasts*, July 12, 2024.

¹¹⁶ Friend, interview; Linebaugh, “Teens are Falling Victim”

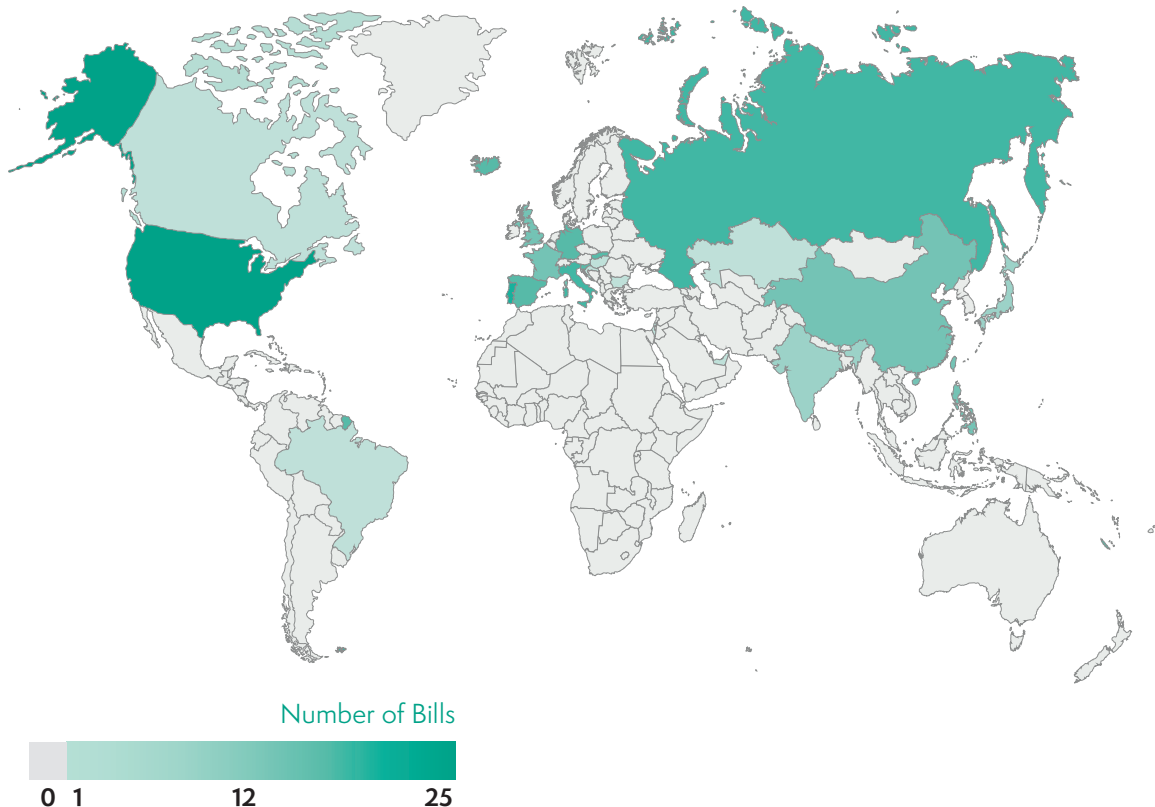
¹¹⁷ Valerie Hudson, “The Right Way to Deal with AI-Generated Child Pornography,” *Deseret News*, June 2, 2024.



II. “Enforcement Can Only Move as Fast as the Speed of Law”¹¹⁸

CURRENT LEGAL STATUS OF AI-GENERATED CSAM

FIGURE 8: AI-RELATED BILLS PASSED INTO LAW BY COUNTRY (2016–2023)¹¹⁹



One hinderance to effectively combatting the rise of AI-generated CSAM is the ability to charge and prosecute perpetrators. Understandably, the legislation surrounding AI-generated CSAM and AI as a broader topic is still in its infancy. As policymakers begin to wrestle with how to respond to the threat of AI-generated CSAM, understanding the current legislative landscape is necessary for a concerted, global response.

Legislating AI

As is often the case with innovative technologies, policymakers prefer to hold off on implementing legal guardrails, due to a lack of understanding or so as not to stymie innovation.¹²⁰ This is evident in the case of AI legislation. Among legislators globally, there are broader tendencies to promote technological and economic advancement ahead of human rights concerns and often reflect society’s general naïveté about the potential for misuse. However, since 2019, AI-related bills have increased by ~150% annually, showing growing interest.

¹¹⁸ Attwood & Bailey, interview.

¹¹⁹ Stanford HAI, AI Index 2024 Annual Report, 2024.

¹²⁰ Cecilia Kang & Adam Satariano, “As A.I. Booms, Lawmakers Struggle to Understand the Technology,” *The New York Times*, March 3, 2023.

A Stanford Internet Observatory mapping indicates that most AI legislation focuses on economics, public finance, and labor.¹²¹ Notably, protecting vulnerable populations is not a key priority among early legislation on AI.

As AI legislation is still developing, national strategies offer insight into how countries are addressing AI. While most countries with AI strategies do not address the risk of AI exacerbating child sexual abuse, three have begun to recognize these dangers. Iceland, for instance, mentions digital abuse and the need to protect vulnerable groups, including children, though it doesn't specifically address AI-generated CSAM. Australia and the United States go further; Australia calls for industry safeguards against AI-generated CSAM, and the United States highlights the need for federal legislation to combat its spread.¹²²

Legislating AI-Generated CSAM

As few countries have explicitly addressed AI-generated CSAM from a legal perspective, determining the global legality of AI-generated CSAM is complex. The legality of CSAM can be categorized broadly as:

1. No specific legal statutes surrounding CSAM:

Just a few countries—namely the Central African Republic, Lesotho, Libya, Moldova, and Syria—have not explicitly made the possession or creation of CSAM illegal. In some of these countries their legal codes may prohibit all sexual material, regardless of the age of the participants, but these countries have not gone to the extent of creating specific legislation around CSAM.

2. Existing legal recourse for CSAM, which could apply to AI-generated CSAM:

Most countries have adopted the international standard of (1) passing legislation specific to CSAM, (2) defining child sexual abuse material, (3) prohibiting technology-facilitated CSAM offenses, and (4) outlawing “simple possession” of

CSAM.¹²³ Depending on the specific wording of the legislation—for example, the inclusion of words like “photorealistic”—some of these countries may be able to prosecute perpetrators for creating, possessing, and/or distributing AI-generated CSAM. However, lacking a specific legal prohibition may mean that it is more difficult to prosecute perpetrators for crimes solely related to AI-generated CSAM, as many of the existing laws have not been tested for this crime. This was the case in March 2024 for U.S. Digital Forensic Examiner David Haddad:

*“From a prosecution perspective, we don't want to be the first ones to prosecute a case that is purely based on AI-generated CSAM and set the precedent. At this point, we have not had a single case that has exclusively considered AI-generated CSAM.”*¹²⁴

Still, other countries have already tested the application of existing CSAM law to AI-generated CSAM, as was the case in Canada. In 2023, a Canadian man was sentenced to over three years in prison for creating “synthetic” CSAM using deepfake technologies. He was additionally charged with and sentenced for possessing countless CSAM files.¹²⁵ Given the difficulty of determining what is AI-generated CSAM, and the fact that most perpetrators still possess both types, some countries may not be pushing as hard for specific AI-generated CSAM legislation. Indeed, countries may already be able to take legal action against perpetrators for crimes related to AI-generated CSAM, but without clarification around the treatment of AI-generated materials, prosecutors and courts leave room for perpetrators to get away with their crimes.

3. Explicit legal recourse for AI-generated CSAM:

Approximately 18% of countries either already have specific AI-generated CSAM legislation or are in the process of passing such legislation.¹²⁶ Primarily within the last two years, countries have begun implementing different means by which to penalize the creation,

¹²¹ Stanford University Human-Centered Artificial Intelligence, “Artificial Intelligence Index Report 2024,” Stanford Internet Observatory, 2024.

¹²² OECD.AI, “OECD AI Incidents Monitor,” OECD.AI Policy Observatory, accessed 30 March and 11 April 2024.

¹²³ International Centre for Missing & Exploited Children, “Child Sexual Abuse Material: Model Legislation & Global Review,” 10th Edition, 2023.

¹²⁴ Haddad, interview.

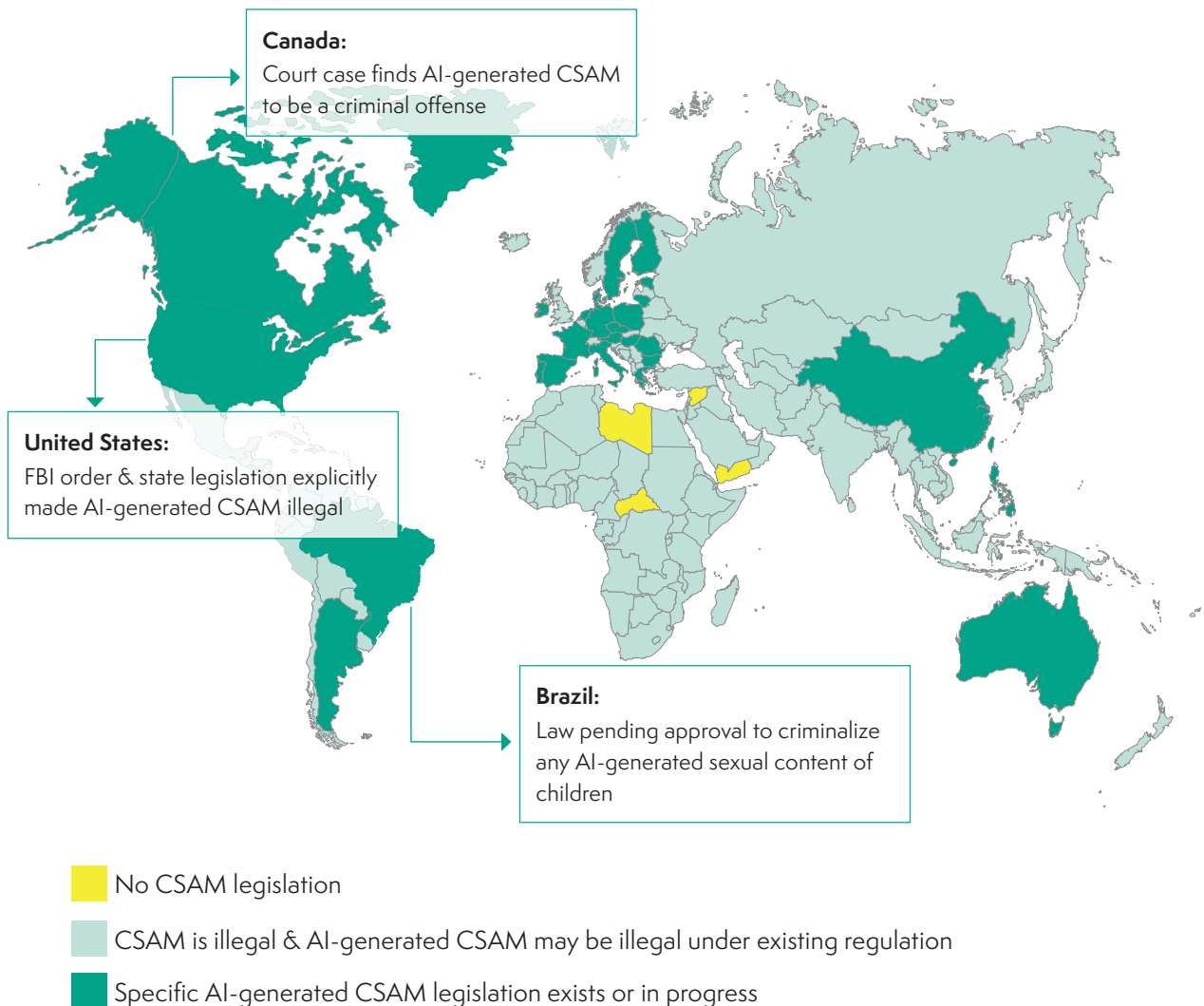
¹²⁵ Jacob Serebrin, “Quebec Man who Created Synthetic, AI-Generated Child Pornography Sentenced to Prison,” *The Canadian Press*, April 26, 2023.

¹²⁶ Data Commons, “World Demographics,” accessed June 2, 2024.

possession, and/or distribution of AI-generated CSAM. In Brazil, for instance, discussions are underway to update the “Child and Adolescent Statute” to include a penalty of four to eight years in prison and a fine for anyone who produces AI-generated CSAM.¹²⁷ In the United States, the FBI released a statement explicitly informing the public that AI-generated CSAM is CSAM.¹²⁸ For member states of the European Union, updates to the existing Child Sexual Abuse Directive (2011/93/EU) that prohibits CSAM,

will likely include a mention of AI-generated CSAM and will require that all member states harmonize the criminalization of AI-generated CSAM.¹²⁹ Furthermore, China has outlawed AI-generated CSAM by banning all “deepfake pornography” and Australia has adopted regulations requiring internet search engines to block AI-generated CSAM.¹³⁰ It remains to be seen which approach to legislation is most effective for combatting the emerging threat.

FIGURE 9: LEGAL STATUS OF AI-GENERATED CSAM WORLDWIDE¹³¹



¹²⁷ Murilo Souza, “Projeto Prevê Até 8 Anos de Prisão para quem Usar Inteligência Artificial para Gerar Conteúdo Sexual com Crianças,” *Camara dos Deputados*, February 27, 2024.

¹²⁸ Federal Bureau of Investigation, “Child Sexual Abuse Material Created by Generative AI and Similar Online Tools is Illegal,” *Public Service Announcement*, March 29, 2024.

¹²⁹ European Commission, “Q&A - The Fight Against Child Sexual Abuse Receives New Impetus with Updated Criminal Law Rules,” *Press Corner*, February 6, 2024.

¹³⁰ Asha Hemrajani, “China’s New Legislation on Deepfakes: Should the Rest of Asia Follow Suit?” *The Diplomat*, March 8, 2023; Josh Butler, “Search Engines Required to Stamp Out AI-Generated Images of Child Abuse Under Australia’s New Code,” *The Guardian*, September 7, 2023.

¹³¹ Stanford HAI, *AI Index 2024 Annual Report*, 2024.

Unique Challenges in Legislating AI-Generated CSAM

Even while many countries appear to be moving in the direction of setting up or adapting protections for AI-generated CSAM, a few unique legislative challenges have arisen.

Existing laws focus primarily on the creation, possession, and distribution of AI-generated CSAM by individual perpetrators. The laws do not focus on the capabilities of generative AI models to create AI-generated CSAM or the content included in the training datasets. Additionally, legislators will need to address the complex legal question of whether creators of AI models can be held accountable for the potential harm to children. Moreover, it remains to be seen how legislators will handle social media and app stores—which are playing a critical role in spreading AI-generated CSAM on mainstream platforms. In Canada, a proposed bill to address online harms would hold social media platforms accountable for addressing harmful content on their platforms, specifically

targeting harmful content that sexually victimizes a child.¹³² These questions of accountability and how to legislate it will need to be answered soon to effectively legislate and prosecute harm perpetuated by AI-generated CSAM.

Moreover, the borderless nature of the crimes related to AI-generated CSAM makes prosecution all the more difficult. For instance, where AI-generated CSAM was viewed can differ from where the model was made, where the content is hosted, and, if there is a real child involved, where the child victim lives. While international cooperation for finding perpetrators is strong, determining which jurisdiction crimes fall under and what is illegal will be challenging if not all countries adopt a similar approach to legislating AI-generated CSAM.

To better understand how the challenges are playing out in practice, we will explore the legal status of AI-generated CSAM in three countries: Japan, the United States, and the United Kingdom, which represent a variety of different responses to this issue.



“We cannot arrest our way out of this problem.”

Simon Bailey, Director of Strategic Engagement, Child Rescue Coalition

¹³² Online Harms Act, C-63, Government of Canada (2024).



Deep Dive: Japan

Despite signing international agreements against AI-generated CSAM, Japan's response has lagged behind other countries. This is arguably related to its past handling of CSAM.¹³³ Japan was criticized by the child safety community in the 1990s for being a hub for child pornography and only criminalized CSAM possession in 2014, the last of the major economies to do so.¹³⁴

Cultural factors may have contributed to Japan's delayed response to addressing CSAM. The concept of "kawaii" (cute) influenced tolerance toward sexualized depictions of children in anime and manga, Japanese styles of animation and graphic novels. These industries have resisted stricter regulations, citing constitutional protections of free expression.¹³⁵ While the 2014 law criminalized possession of CSAM, this ban did not extend to animated content or manga.¹³⁶ Still, the legislation faced significant resistance from artists and publishers, concerned about artistic freedom.¹³⁷

Given this, regulating AI-generated CSAM is particularly difficult in Japan. Under the 2014 law, computer-generated content is only considered unlawful if it intentionally resembles actual children in appearance.¹³⁸ Therefore, AI can produce hyper-realistic images that exploit legal loopholes and fall outside current child protection laws.¹³⁹ In 2016, for instance, a Tokyo court found a man guilty of violating the law for creating and offering for sale computer-generated images of a young naked girl; this case only resulted in conviction because the AI-generated images closely matched a real child.¹⁴⁰

AI platforms and other tech platforms have been exploiting this gap in the legislative protections by producing and distributing highly realistic images of child sexual abuse on such Japanese sites as

Pixiv, which are then accessible globally.¹⁴¹ Efforts by children's rights activists to ban explicit depictions in manga and anime have so far been unsuccessful and attempts to broaden the scope of laws to include AI-generated materials face resistance from powerful industries and constitutional concerns, hindering legislative progress.¹⁴² For the rest of the world, Japan's current approach to AI-generated CSAM may pose significant risks if AI-generated CSAM can be hosted there.

Despite this, Japan is exploring ways to address AI-generated CSAM. Amidst strong criticism from the manga and anime industry, the Hiroshima AI Process, launched during Japan's G7 chairmanship in 2023, aims to enhance online safety.¹⁴³ Joined by over 49 countries, it addresses intellectual property, disinformation, and AI governance. The initiative collaborates with tech companies to combat child sexual exploitation and ensure children's online safety and privacy. It includes guiding principles and a code of conduct for AI actors, emphasizing transparency and digital literacy. Japan aims to lead global AI governance through the AI Safety Institute and the Tokyo Center of the Global Partnership on AI.

Japan's manga and anime industries' stance on copyright and AI might help close the AI-generated CSAM loophole. In January 2024, the Japan Agency for Cultural Affairs released a draft "Approach to AI and Copyright" to clarify the use of copyrighted materials in AI applications. This bill follows concerns from creators regarding inadequate copyright protection under Japan's 2019 provisions, which permitted extensive use of copyrighted works, including for commercial purposes and from illegal sources. The revised 2019 Copyright Act made Japan highly AI-friendly by permitting the use of copyrighted works for AI model training.¹⁴⁴ This push by manga and anime creators for copyright protection contrasts the

¹³³ Tommy Chriyst, "Reflecting on the 2024 AI Safety Summit and What it Means," *Linkdln*, May 22, 2024; Government of the United Kingdom, "Home Secretary Urges Meta to Protect Children from Sexual Abuse," *Gov.uk*, September 20, 2023.

¹³⁴ Victoria Macchi, "Japan Outlaws Owning Child Pornography," *Learning English*, July 15, 2014; Yomiuri Shimbun, "Japan Lags in Regulating AI-Generated Child Porn as Loophole in Existing Law Gets Exploited," *The Japan News*, November 14, 2023.

¹³⁵ Julian Ryall, "Realistic AI-Generated Child Porn in Japan Sparks Debate on Legal Loophole and 'Kawaii' Culture," *Everand*, November 17, 2023.

¹³⁶ BBC, "Japan Bans Child Pornography Possession," *BBC*, June 18, 2014.

¹³⁷ BBC, "Japan Bans Possession."

¹³⁸ Shimbun, "Japan Lags in Regulating."

¹³⁹ Christian Martini Grimaldi, "Pedophilia Is Not Taboo Enough in Japan," *UCA News*, June 3, 2024.

¹⁴⁰ Mary-Ann Russon, "Japan: Artist Guilty of Creating Computer-Generated VR Child Pornography of an Actual Girl," *International Business Times*, March 18, 2016.

¹⁴¹ Ryall, "Legal Loophole and 'Kawaii' Culture."

¹⁴² Ryall, "Legal Loophole and 'Kawaii' Culture."

¹⁴³ GIP Digital Watch Observatory, "Hiroshima AI Process: G7's Effort to Tackle Challenges of AI Technology," *Dig.Watch*, May 21, 2023.

¹⁴⁴ Scott Warren & Joseph Grasser, "Japan's New Draft Guidelines on AI and Copyright: Is it Really OK to Train AI Using Pirated Materials?" *Privacy World*, March 12, 2024.

industry's acceptance of CSAM being fed to AI training data and their defense of AI-generated CSAM for "artistic" purposes. New in-progress regulations, such as penal codes for foundation model developers like OpenAI, may help close the door on AI-generated CSAM in Japan.¹⁴⁵



Deep Dive: United States

On March 29, 2024, the FBI issued a public service announcement warning that CSAM created with generative AI is illegal under federal law, which prohibits the production, advertisement, sale, possession, and "access with intent to view" of any CSAM, including "realistic computer-generated images."¹⁴⁶ Federal legislative bodies are working to codify these protections in law. A few examples include:

- **President Biden's October 2023 Executive Order** called for identifying existing and potential development of standards, tools, methods, and practices for preventing generative AI from producing CSAM or producing non-consensual intimate imagery of real individuals.¹⁴⁷
- The Senate's 2024 proposed **Take It Down Act** and the **Defiance Act** endeavor to curb non-consensual intimate imagery and any "digital forgeries."¹⁴⁸
- The **House of Representatives** has introduced bills to update the wording of existing legislation to ensure creating and distributing AI-generated abuse and exploitation material is clearly criminalized under federal law.¹⁴⁹
- **The EARN IT Act**, reintroduced in the Senate in 2023, aims to limit the liability protections of "interactive computer services providers (e.g., Facebook and Twitter)" related to CSAM.¹⁵⁰

As these bills go through the legislative process, it is clear that the U.S. government recognizes the importance of criminalizing the possession and distribution of AI-generated CSAM. However, there is not yet a push to legally restrict the underlying technology.

Alongside federal legislation, U.S. states are deploying various methods to criminalize AI-generated CSAM:

- **Updating definitions:** Florida, South Dakota, and Washington have updated their definitions of CSAM (or child pornography) to include AI-generated and/or deepfake technology.¹⁵¹ Washington's "Fabricated Intimate or Sexually Explicit Images" state law updated definitions for "photograph," "visual or printed matter," and "sexually explicit conduct;" "a fabricated depiction of an identifiable minor" will hopefully cover future criminal innovations.¹⁵² By expanding definitions, lawmakers hope to combat current and future crimes.
- **Amending existing laws:** Indiana, New York, and Virginia have addressed "deepfakes" in their existing prohibitions on "revenge porn."¹⁵³ Georgia and Hawaii have updated privacy laws to include "deepfake porn."¹⁵⁴
- **Drafting new laws:** A 2023 Texas law specifically targeted sexually explicit "deepfakes" distributed without the subject's consent.¹⁵⁵ As of June 2024, a California bill criminalizing the "creation, distribution, and possession" of AI-generated CSAM awaits a vote in the State Senate.¹⁵⁶

¹⁴⁵ Warren & Grasser, "Japan's New Draft."

¹⁴⁶ FBI, "Child Sexual Abuse Material Is Illegal."

¹⁴⁷ The White House, "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," 2023.

¹⁴⁸ Alexandria Ocasio-Cortez, "Rep. Ocasio-Cortez Leads Bipartisan, Bicameral Introduction of DEFIANCE Act to Combat Use of Non-Consensual, Sexually-Explicit 'Deepfake' Media," Press Release, March 7, 2024; The TAKE IT DOWN Act, S.4569, 118th Cong. (2024).

¹⁴⁹ Sexual Exploitation and Other Abuse of Children, U.S. Code 18 (2024), §§ 2251-2260A; Sexual Exploitation and Other Abuse of Children, U.S. Code 18 (2024), §§ 1460-1470; Nate King, interview by Value for Good, June 25, 2024.

¹⁵⁰ Eliminating Abusive and Rampant Neglect of Interactive Technologies (EARN IT) Act of 2023, S.1207, 118th Cong. (2023).

¹⁵¹ Florida Senate, Sexually Explicit Material, SB.1798 (2022); South Dakota Legislature, An Act to Revise Provisions Related to the Possession, Distribution, and Manufacture of Child Pornography, SB.79 (2024); Washington State Legislature, Fabricated Intimate or Sexually Explicit Images, H.1999 (2024).

¹⁵² Washington State Legislature, Fabricated Intimate.

¹⁵³ New York State Senate, Unlawful Dissemination or Publication of Intimate Images Created by Digitization, S.1042A (2023-24); Indiana General Assembly, Sexual Offenses, H.1047 (2024); New York State Senate, Unlawful Dissemination or Publication; Virginia General Assembly, Unlawful Dissemination or Sale of Images of Another, H.2678 (2019).

¹⁵⁴ Georgia General Assembly, Invasion of Privacy; Prohibition Against the Transmission of Photography Depicting Nudity; Include Falsely Created Videographic or Still Images, SB.337 (2020); Hawaii State Legislature, A Bill for an Act, S.309 (2021).

¹⁵⁵ Texas State Legislature, Relating to the Unlawful Production or Distribution of Sexually Explicit Videos Using Deep Fake Technology; Creating a Criminal Offense, SB.1361 (2023).

¹⁵⁶ Marc Berman, "Bill to Protect Children from AI Enabled Sexual Exploitation Passes Assembly," Press Release, May 23, 2024.

- **Setting precedents:** Although court cases involving AI-generated CSAM in the U.S. are still few, recent decisions are setting important precedents. As of June 2024, a Wisconsin man awaits trial for producing, distributing, and possessing AI-generated images of children engaged in explicit acts.¹⁵⁷ Notable prior cases include a North Carolina child psychiatrist sentenced to 40 years in November 2023 for child sexual exploitation and using AI to create CSAM and a Pennsylvania man given 15 years for possessing CSAM depicting child celebrities.¹⁵⁸ These cases will likely influence how both federal and state governments handle AI-generated CSAM in the future.

Despite efforts to future-proof legislation, these states may fall short of creating a robust framework that protects future generations. These quick fixes often focus on individual actors rather than those who create models or host content. Some political strategists intentionally avoid blaming creators to prevent stifling innovation.¹⁵⁹ While this approach prioritizes innovation, it remains to be seen whether a more dedicated effort is needed to combat the issue effectively.



Deep Dive: United Kingdom

The United Kingdom is at the forefront of legislating AI-generated CSAM, not only addressing perpetrators who possess and distribute this material, but also endeavoring to hold social media companies and search services responsible for surfacing this content.

Three key laws aim to protect children from abuse through AI-generated CSAM. The first two are existing statutes that can be used to prosecute AI-generated CSAM. Depending on the specific AI-generated CSAM at hand, prosecutors can try

perpetrators under the auspices of either The Protection of Children (PoC) Act of 1978 or The Coroners and Justice Act of 2009.¹⁶⁰ Under the PoC Act (which was amended by the Criminal Justice and Public Order Act of 1994), the main criterion for deeming an image criminal is that it “appears to be a photograph.” It is not necessary to prove whether the image is AI-generated; it only needs to resemble a photograph and be an indecent image of a child to be prosecutable under the Act. Part of the job of law enforcement therefore becomes determining whether an image is realistic enough to be an indecent pseudo-photograph.

According to The Coroners and Justice Act of 2009, an image must only meet the threshold of “prohibited image of a child” to be criminal—a classification that under U.K. law results in lighter sentences for offenders. This act primarily refers to non-photographic content such as cartoons, drawings, and animations, but could be applied to AI-generated CSAM as well. Courts do not need to determine if an image is “real” or AI-generated to prosecute a perpetrator under the existing laws. Courts only need to assess if the image meets the necessary legal criterion of appearing to be a photograph or being a prohibited image of a child. This approach ensures that AI-generated images can be prosecuted under current legislation.¹⁶¹ Moreover, the U.K. government made the creation of sexually explicit deepfake images a new offense in April 2024, ensuring that a perpetrator can be charged for creating and sharing such images.¹⁶²

As the first cases involving AI-generated CSAM make their way through the U.K. court system, some still fear that prosecution of AI-generated CSAM could result in lesser sentences because they cannot identify a child who was harmed in the creation of such images.¹⁶³ A U.K. man, convicted in April 2024 for creating over 1,000 indecent images of children with AI, has been ordered to pay a £200 fine.¹⁶⁴ This decision includes a

¹⁵⁷ Office of Public Affairs, U.S. Department of Justice, “Man Arrested for Producing, Distributing, and Possessing AI-Generated Images of Minors Engaged in Sexually Explicit Conduct,” Press Release, May 20, 2024.

¹⁵⁸ United States Attorney’s Office, Western District of North Carolina, “Charlotte Child Psychiatrist Is Sentenced to 40 Years In Prison For Sexual Exploitation of A Minor And Using Artificial Intelligence To Create Child Pornography Images Of Minors,” Press Release, November 8, 2023; Office of Public Affairs, U.S. Department of Justice, “Recidivist Sex Offender Sentenced for Possessing Deepfake Child Sexual Abuse Material,” Press Release, May 1, 2024.

¹⁵⁹ Madyson Fitzgerald, “States Race to Restrict Deepfake Porn as it Becomes Easier to Create,” *Utah News Dispatch*, April 12, 2024.

¹⁶⁰ Internet Watch Foundation, “How AI Is Being Abused.”

¹⁶¹ Shinyshiny, “Do AI Images Count as Indecent Images?,” Shinyshiny.tv., May 15, 2024.

¹⁶² Government of the United Kingdom, “Government Cracks Down on ‘Deepfakes’ Creation,” Gov.uk, Ministry of Justice, April 16, 2024.

¹⁶³ IWF, “Child Sexual Abuse Imagery.”

¹⁶⁴ Shanti Das, “Sex Offender Banned from Using AI Tools in Landmark UK Case,” *The Guardian*, April 21, 2024.

prohibition on accessing generative AI tools without police permission as part of a sexual harm prevention order. The ban specifically targets text-to-image generators and “nudifying” websites, including the Stable Diffusion software.¹⁶⁵ While it is not clear how many cases involving AI-generated CSAM have already made their way to the U.K. courts—*The Guardian* tallied at least nine in April 2024—these initial cases are important to set the precedent for how courts view CSAM.¹⁶⁶

The third law that makes up the United Kingdom’s legal response to AI-generated CSAM is the 2023 Online Safety Act. The act focuses on child protection and removing illegal content from social media companies and search services.¹⁶⁷ Under this new law, social media platforms must remove illegal content and prevent children from accessing harmful materials, such as pornography. Non-compliance could result in fines for companies by Ofcom, the U.K. broadcasting regulator, of up to £18 million (\$22.3 million) or 10% of global turnover.¹⁶⁸ This piece of legislation, opposed by some tech companies over fears that they will have to break their end-to-end encryption, is a first in holding social media companies and search services responsible for what is available on their platforms.

In addition to its legal framework, the United Kingdom has been vocal in raising the issue of AI-generated CSAM on the world’s stage, including:

- In September 2023, the United Kingdom and the United States issued a **joint statement**, pledging to “combat” AI-generated CSAM.¹⁶⁹ The leaders of these two countries reaffirmed their commitment to battle child sexual abuse globally, recognized the challenges posed by evolving technology like generative AI, pledged to collaborate on innovative solutions and support organizations like NCMEC, and urged international cooperation to prevent online exploitation and bring perpetrators to justice.¹⁷⁰

- The U.K. government endorsed the **Bletchley Declaration** in November 2023, a global agreement that states a united commitment to manage risks associated with advanced AI models, ensuring their safe and responsible development and deployment.¹⁷¹ Signatories agree to identify AI safety risks through scientific research and create risk-based policies to address them. For the private sector, the Declaration highlights the global opportunities AI presents and advocates for a governance and regulatory approach that balances innovation with safety.

The United Kingdom will undoubtedly need to continue testing its legal framework against developments in generative AI. The current legislative landscape reflects an awareness of the threat posed by AI-generated CSAM. However, the sentences given to perpetrators suggest that the courts may be slower in recognizing the harm caused by AI-generated CSAM.

¹⁶⁵ Shanti Das, “Sex Offender Banned.”

¹⁶⁶ Shanti Das, “Sex Offender Banned.”

¹⁶⁷ U.K. Department for Science, Innovation & Technology, “What the Online Safety Act Does,” Gov.uk, May 8, 2024.

¹⁶⁸ Paul Sandle, “UK’s Online Safety Bill Finally Passed by Parliament,” *Reuters*, September 19, 2023.

¹⁶⁹ Government of the United Kingdom, “UK and US Pledge to Combat AI-Generated Images of Child Abuse,” Gov.uk, September 27, 2023.

¹⁷⁰ Government of the United Kingdom, “A Joint Statement from the United States and the United Kingdom on Combatting Child Sexual Abuse and Exploitation,” Gov.uk, September 27, 2023.

¹⁷¹ Government of the United Kingdom, “Countries Agree to Safe and Responsible Development of Frontier AI in Landmark Bletchley Declaration,” Gov.uk, November 2023.

III. “Even Offenders Won’t Be Able to Differentiate”¹⁷²

DETECTING AI-GENERATED CSAM

To attempt to counter AI-generated CSAM, both tech companies and law enforcement need to be able to detect its presence, whether that be on a company’s platform or during a police investigation. Detection is becoming increasingly challenging as the technology behind AI-generated CSAM improves and the volume of content grows. While tech companies and law enforcement agencies work towards solutions, the continually evolving maturity of AI technology makes progress difficult. This section covers the state of detection capabilities and major hurdles for law enforcement and tech platforms.

Law Enforcement and Detection

While generative AI has advanced at a rapid pace, the technology available to law enforcement to detect abuses of AI, including CSAM, has not. Detection still relies primarily on manual techniques. When officers obtain perpetrators’ devices, they often contain hundreds of thousands of images to work through. As Officer Frend explains it, police work used to be focused on categorizing images based on the severity of abuse, but with the encroachment of generative AI, analysts must now also scrutinize images to determine if they are AI-generated.¹⁷³ While all CSAM, including that which is AI-generated, is illegal in Officer Frend’s jurisdiction, this differentiation is still necessary to determine if there are new victims in need of safeguarding.

To determine whether an image has been manipulated or created with AI, an analyst will look for tells. Common tells that indicate AI manipulation, at least as of late 2023, are the number of fingers on hands or the absence of blinking in videos. These tells were relatively easy to spot in 2023, though still required more intense analysis from law enforcement. But since then, the techno-

logy has improved dramatically and the tells have been addressed by modelers, hindering law enforcement’s efficacy and speed.¹⁷⁴ Data from the survey of global law enforcement conducted for this report in spring 2024 indicates that this hinders law enforcement’s response; while ~30% of law enforcement agents reported that identifying AI-generated material was possible, ~70% found it difficult.¹⁷⁵ This tension could be attributed to some law enforcement agents not realizing how quickly generative AI has evolved.

In addition to slowing down the identification of victims, the psychological impact of analyzing CSAM increases the longer an officer must examine the images.¹⁷⁶ While most officers working on this crime receive specialized training, this training primarily concerns categorizing images based on severity and does not include how to distinguish AI-generated content.¹⁷⁷ In fact, 79% of global law enforcement agents surveyed in 2024, reported that they did not have enough training at their disposal to discern AI-generated CSAM.¹⁷⁸

Often the investigative burden of this crime, including the analysis and classification of images, falls to the most under-resourced officers. The majority of CyberTips originate in high-income countries, but often end up on the desks of law enforcement officers in low- and middle-income countries.¹⁷⁹ Officers in those countries are likely to have less training and fewer resources, meaning that even if CyberTips come in and CSAM is detected, the investigation burden will land on these officers.

As the obvious tells of AI-generated CSAM fall away, there are increasing calls for tools that can reliably detect this material; however, no such tools exist. Officers and AI-detection experts note that there are various tools that can sometimes tell what AI-modeling software was used to create an image or, if there is a source image available,

¹⁷² Attwood & Bailey, interview.

¹⁷³ Frend, interview.

¹⁷⁴ Attwood & Bailey, interview.

¹⁷⁵ Value for Good, survey of global law enforcement, spring 2024.

¹⁷⁶ Frend, interview.

¹⁷⁷ Frend, interview.

¹⁷⁸ Value for Good, survey of global law enforcement.

¹⁷⁹ Grossman et al., “The Strengths and Weaknesses”.

may be able to find what AI-generated CSAM has been created from it.¹⁸⁰ Officers are also able to investigate the metadata of digital images, which can show if an image has been manipulated by AI. However, in the lifecycle of an image, these data can be wiped, for example when sent through a communication channel such as WhatsApp.

Even if reliable tech were to exist, knowing whether an AI tool was used may still not help with triaging and safeguarding, as AI has the power to disguise the victim by changing identifying features or creating composites of victims. Still, the promise of a method to distinguish AI-generated CSAM is hoped for in the law enforcement community.

Tech Companies and Detection

Tech companies are also searching for the key to detect AI-generated content, including CSAM, on their platforms. Tech platforms, such as Meta and TikTok, have a variety of reasons for not wanting malicious AI-generated content on their sites. Firstly, the content can open these companies up to government action, such as fines and lawsuits, especially if their platforms host content that incites violence or destabilizes governments. Additionally, too much explicit content could drive their desired consumers away, creating reputational risk, which threatens advertising dollars.

In response, many companies have signed on to safety-by-design principles, such as those set forth by Thorn and All Tech Is Human in spring 2024.¹⁸¹ Even with these drivers for developing detection methods, the industry remains at a loss for how to do so effectively.

Industry leaders are engaging in defense as they attempt to detect this content and keep it off their platforms. Many social media companies, including Meta, employ AI for the detection and

reporting of suspicious material, which is then moved on to human moderators before being removed and/or turned over to law enforcement.¹⁸² The reliance on human moderators is necessary, but also harmful, as moderators are exposed to explicit images constantly. The hope is that better AI would eliminate the need for human content moderators. Meta is currently building tools that should be able to identify images made using common AI platforms, like OpenAI and Midjourney, which will then be able to label AI-generated content on their sites.¹⁸³ Recently, Instagram has begun to include markers that say “Made with AI” on posts, if that metadata is available.¹⁸⁴

Although AI-generated content on social media platforms is an international issue, U.S. law has a major impact on the spread of CSAM on online platforms. The majority of major platforms—Google, Meta, X—are based in the United States, giving the United States more responsibility in addressing the issue. Under U.S. federal law, ESPs must report both CSAM and AI-generated CSAM to NCMEC, but they are not required to actively look for CSAM on their platforms.¹⁸⁵

While platforms, including Google, have adopted such safety-by-design principles as the “Priority Flagging Program”—a partnership between Google and third parties who flag potentially violative content—the focus is often not on CSAM when it comes to identifying AI-generated content.¹⁸⁶ Due to the political ramifications of spreading misinformation about elections or the government, tech platforms appear to prioritize addressing the spread of such material.¹⁸⁷ The potential political ramifications are in fact what pushed OpenAI to launch its “Deepfake Detector;” they posited in May 2024 that it can detect ~99% of AI-generated images created by its own image generator—DALL-E—but is not usable on other image generators, like Midjourney or Stability AI.¹⁸⁸ As tech companies search

¹⁸⁰ Frend, interview; Attwood & Bailey, interview; Oldroyd, interview.

¹⁸¹ Emma Woollacott, “Tech Firms Pledge To Eliminate AI-Generated CSAM,” *Forbes*, April 24, 2024.

¹⁸² Katie McQue, “Revealed: US Police Prevented from Viewing Many Online Child Sexual Abuse Reports, Lawyers Say,” *The Guardian*, January 17, 2024.

¹⁸³ Levine, “Stable Diffusion 1.5 Was Trained.”

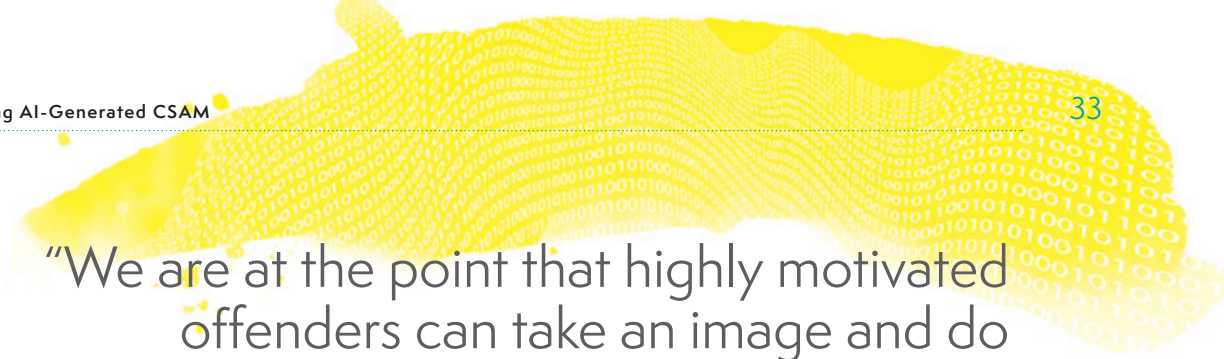
¹⁸⁴ Sourabh Singh, “Instagram Identifying Photographs that Have Been Modified with AI,” *United Business Journal*, June 20, 2024.

¹⁸⁵ Oremus, “AI Is About to Make.”

¹⁸⁶ Susan Jasper, “An Update on our Child Safety Efforts and Commitments,” Google Blog, Safety & Security, April 23, 2024.

¹⁸⁷ Oldroyd, interview.

¹⁸⁸ Alex Popken, “We’re Unprepared for the Threat GenAI on Instagram, Facebook, and WhatsApp Poses to Kids,” *Fast Company*, June 7, 2024; Cade Metz & Tiffany Hsu, “OpenAI Releases ‘Deepfake’ Detector to Disinformation Researchers,” *The New York Times*, May 2024.



“We are at the point that highly motivated offenders can take an image and do whatever they want with it.”

Simon Bailey, Director of Strategic Engagement, Child Rescue Coalition

for solutions to problems their platforms created, they are placing their trust in the development of new technologies—a cycle that could prove to be counterproductive.

Many tech solutions companies are also working on detection, with some reporting that they already can detect AI-generated materials, including CSAM.¹⁸⁹ However, others, such as Semantics21, a tech solutions company developing tools for law enforcement to detect CSAM, urge people to be more skeptical of such claims, arguing that if a solution were available, all major platforms would have already adopted it.¹⁹⁰ Tom Oldroyd from Semantics21 notes that if metadata are wiped, the detectors have a relatively low success rate.¹⁹¹ According to Oldroyd, this is in large part due to the rapid advancements in AI; while a detector may work well for a few weeks, an advancement in AI could reduce its efficacy.¹⁹²

Currently, available detection tools are able to identify whether an image has been tampered with. They may not be able to tell how exactly AI was used, but their detection methods do indicate what may have been adjusted or if the image is 100% original.¹⁹³ These detectors can be thought of much in the same way as calculators or x-ray machines. They can provide an analyst with data not available to the naked eye, but for a meaningful conclusion or to ensure these data are interpreted correctly, the analyst must use their judgment and experience.¹⁹⁴

The Future of Detection for AI-Generated CSAM

As generative AI improves, the likelihood that a silver bullet detection technology is developed is low. In fact, it becomes increasingly likely that perpetrators themselves will not be able to differentiate AI-generated content.¹⁹⁵ To effectively combat AI-generated CSAM, law enforcement reports the greatest need for effective tools and training.¹⁹⁶ In terms of specific tools, law enforcement envisions AI-powered tools that support them in triaging which images to focus on first and “heat maps” that identify which part of an image could be real and which is AI-generated.¹⁹⁷ Notably, these proposed tools do not remove the human element from the analysis, rather they emphasize the importance of a technology that could ease the burden on law enforcement.

Many open questions remain about how these interventions could be developed to meet future needs. For instance, a tech solution will have to address the possibility that perpetrators are exploiting AI tells and hiding a real child behind an AI-manipulated image. A tech solution will also have to grapple with the ethical question of whether it can be trained on CSAM in its development. Until a reliable tool is created, law enforcement must develop intermediate solutions that allow them to safeguard children and take down large-scale perpetrators along the way.

189 Oldroyd, interview.

190 Nitasha Tiku & Tatum Hunter, “Fooled by AI? These Firms Sell Deepfake Detection That’s ‘REAL 100%,’” *The Washington Post*, May 12, 2024; Oldroyd, interview.

191 Oldroyd, interview.

192 Oldroyd, interview.

193 Martino Jerio & Marco Fontani, interview by Value for Good, June 11, 2024.

194 Jerio & Fontani, interview.

195 Attwood & Bailey, interview.

196 Value for Good, survey of global law enforcement.

197 Frend, interview.

IV. “We’re Behind the Curve”¹⁹⁹

HOW ARE PRIVATE COMPANIES, GOVERNMENTS, LAW ENFORCEMENT, AND CAREGIVERS RESPONDING?

Overcoming the challenges and harms presented by AI-generated CSAM will require a collaborative and coordinated effort from the private sector, governments, law enforcement, and caregivers, each with a key role to play. This section will outline how each stakeholder group is responding to the rise in AI-generated CSAM.



Private Sector

Most tech companies, whether they are developing AI models or a new social network, do not set out to design systems that facilitate the abuse of children.¹⁹⁹ As AI has developed and the public has become more aware of AI’s capabilities, companies that develop and host AI-generated content have reacted in various ways to try to stem the tide of AI-generated CSAM. These reactive measures vary in their efficacy and the piecemeal nature of the response indicates that companies are unsure about how to balance safety, profits, and privacy. A few notable measures will be detailed below:

AI Developers

Google: Safety-by-design principles

Google has adopted AI safety-by-design principles, announcing that they are “proactively implementing guardrails for child safety risks.”²⁰⁰ Their principles that address AI-generated CSAM, as published on their website, include:

- 1. Training datasets:** Integrate hash-matching²⁰¹ and child safety classifiers to remove CSAM as well as other exploitative and illegal content from training datasets.

- 2. Identifying child sexual abuse and exploitation (CSAE) seeking prompts:** Utilize machine learning to identify CSAE-seeking prompts and block them from producing outputs that may exploit or sexualize children.
- 3. Adversarial testing:** Conduct adversarial child safety testing—the process of “proactively trying to “break” an application by providing it with data most likely to elicit problematic output”—across text, image, video and audio for potential risks and violations.
- 4. Engaging experts:** Partner with expert third parties who flag potentially violative content, including for child safety, for review.

Google also publicizes its commitment to flagging content that may indicate a child is in “active danger” and reports the instance to NCMEC.²⁰² More companies, including Amazon, Anthropic, Civitai, Microsoft, OpenAI, and Stability AI, have signed on to similar safety-by-design principles promoted by Thorn and All Tech Is Human, hoping to push for industry-wide change.²⁰³

Other AI developers engage in adversarial testing, or red teaming, of their AI models as well. Though due to regulations regarding AI-generated CSAM, they are often unable to test the models to the extent they would like.²⁰⁴

OpenAI and other AI platforms: Prompt blocking

One technique OpenAI and other AI platforms have adopted to control what users do with AI models is to implement prompt blocking, in which specific topics are blocked by the model. For

¹⁹⁸ Haddad, interview.

¹⁹⁹ Attwood & Bailey, interview.

²⁰⁰ Jasper, “An Update on our Child Safety Efforts.”

²⁰¹ Hash-matching is used in the child protection ecosystem to match CSAM by converting known CSAM into an algorithm and comparing hash lists without looking at the CSAM content itself. Thorn, “How Hashing and Matching Can Help Prevent Revictimization,” August 24, 2023.

²⁰² Jasper, “An Update on our Child Safety Efforts.”

²⁰³ Woollacott, “Tech Firms Pledge.”

²⁰⁴ U.S.-based tech company, interview by Value for Good, August 30, 2024.

example, prompting OpenAI's ChatGPT for anything related to CSAM results in a message reading, "This content may violate our usage policies" and immediate removal of the prompt.²⁰⁵ Savvy users are able to avoid this block by asking the counterfactual of a restricted question. In the 2024 Wisconsin, United States case, the accused was shown to use "negative prompts" that directed Stable Diffusion on what not to include in the AI-generated CSAM, thus avoiding the prompt block.²⁰⁶ While prompt blocking may inhibit some users, it seems to do little to tackle the more egregious perpetrators.

Stability AI: Deletion of not safe for work (NSFW) content

Following the discovery of Stability AI's training set containing CSAM, Stability AI has tried to sanitize its recent model by excluding all explicit content. However, images generated by the latest Stable Diffusion model have serious defects, including multiple arms and no heads.²⁰⁷ Stability AI will have to address how to achieve realistic images of humans, without exploiting sexual images, especially of children.

Internet and Social Media Platforms

Apple: Client-Side Scanning

Apple, and other companies, have considered the idea of deploying client-side scanning, or systems that scan message contents for similarities to a database of objectionable content before the message is sent. This proposal, which would likely drastically stop the spread of AI-generated CSAM, was abandoned by Apple in 2021 due to privacy concerns.²⁰⁸ Opponents fear that client-side scanning would be the start of a wave of content scanning, in which companies would scan by default.

Meta: Financial backing for NCMEC's Take It Down initiative

In early 2023, NCMEC launched a new platform—Take It Down—which aims to support children with

nude content of themselves on the internet to get the material removed. This platform is funded by Meta, showing one way in which Meta is attempting to combat the proliferation of online CSAM. However, Meta and other platforms are balancing competing priorities, including preserving user's right to privacy through encryption. End-to-end encryption, which Meta has implemented on WhatsApp in 2016 and has recently decided to implement on Facebook and Instagram, gets around the goal of Take It Down, as Meta cannot read any content being sent via encrypted messaging. As of September 2023, U.K. law enforcement was arresting 800 perpetrators and protecting up to 1,200 children monthly based on data from social media, including from Meta; the National Crime Agency warned that the recent encryption of Meta products would leave many criminals undetected.²⁰⁹ This conflict of interest is one way in which privacy, profit, and safeguarding children are difficult to unite.

TikTok: Synthetic media policy

TikTok instituted a "synthetic media policy" in 2023, prohibiting all AI-generated content that contains "realistic-appearing people" who seem to be under the age of 18.²¹⁰ They also require any AI-generated content to be labelled as such. How well these policies are enforced is unclear.

X: Implementation of Thorn's Safer AI model

X has recently implemented Thorn's Safer AI model, which X claims is able to seamlessly block text-based online child sexual exploitation and abuse material through proactive detection, deletion, and reporting.²¹¹ The impact and efficacy of this model have not been made public. At the same time, however, the European Commission has raised concerns about X's diminishing content moderation resources. The Safer AI model may help X get on top of the CSAM proliferating on its platforms but may not be enough to meet regulatory standards.

²⁰⁵ Value for Good analysis, 2024.

²⁰⁶ Vallari Sanzgiri, "Here's What the Wisconsin CSAM Incident Tells Us About Prompt-Blocking and E2EE," Medianama, June 11, 2024.

²⁰⁷ Benj Edwards, "New Stable Diffusion 3 Release Excels at AI-Generated Body Horror," ARS Technica, June 12, 2024.

²⁰⁸ Natascha Lomas, "Google's Call-Scanning AI Could Dial Up Censorship by Default, Privacy Experts Warn," Yahoo! Lifestyle, May 15, 2024.

²⁰⁹ Government of the United Kingdom, "Home Secretary Urges Meta to Protect Children."

²¹⁰ TikTok, "Edited Media and AI-Generated Content (AIGC)," TikTok Guidelines, May 17, 2024.

²¹¹ Jason Nelson, "Twitter Touts 'Seamless' Blocking of Child Abuse Content as Elon Musk Faces Increased EU Scrutiny," *Emerge*, May 9, 2024.

Law Enforcement

Law enforcement responses to AI-generated CSAM vary by country and capability. Data from the survey of global law enforcement conducted for this report show how the threat of AI-generated CSAM manifests differently across four countries, namely Canada, the United Kingdom, Portugal, and Nigeria.

The survey results highlight the following takeaways, as presented in figure 10:

- In Canada, law enforcement reports seeing the widest range of generative AI content at the highest rate, from “nudify” tools to AI-generated CSAM of famous individuals, to new modes of sextortion.
- In the United Kingdom, law enforcement sees primarily AI-generated CSAM depicting famous individuals and very few cases of sextortion.
- In Portugal, “nudify” apps are the most prevalent, with law enforcement reporting few interactions with online platforms that distribute AI-generated CSAM and no experience with online platforms refusing to remove AI-generated CSAM.
- Nigeria’s law enforcement has seen financial crimes related to AI-generated CSAM most frequently, which aligns with reports of many sextortion schemes coming out of Nigeria.

How AI-generated CSAM comes across law enforcement’s desk impacts their response as well. Portuguese law enforcement, which has primarily encountered “nudify” apps, reports that it is easy to distinguish AI-generated CSAM. Meanwhile, the United Kingdom, Canada, and Nigeria find it difficult to distinguish which CSAM is AI-generated. As noted earlier, it is unclear whether this reported ease in detecting AI-generated CSAM is because the images being produced are less sophisticated in some jurisdictions or because law enforcement does not yet realize that CSAM they are viewing is AI-generated.

The process of distinguishing AI-generated content remains important for a significant portion of the law enforcement community. Over 60% of law

enforcement surveyed in these four countries still prioritize combatting CSAM, indicating that the process of determining whether a “real” child is in a situation of harm remains critical for law enforcement processes.

In the United Kingdom, they still use traditional techniques, like looking for AI tells in the images.

For police officers, their priority remains making sure they are seeing “first-generation images,” or “net new” CSAM. If this remains the priority, Phil Attwood and Simon Bailey from the Child Rescue Coalition predict that the “challenge of the decade” will be cataloging the sheer volume of material law enforcement has to wade through to find children in harmful situations.²¹²

Law enforcement’s response is highly dependent on the legal framework of their jurisdiction. If legislation empowers them to go after perpetrators who possess AI-generated CSAM, then they will, but if the legal framework does not consider AI-generated CSAM critical, they will have to spend an inordinate amount of time categorizing and proving that images are “real.” In terms of training and tools, law enforcement is clamoring for information on how to understand and identify AI-generated CSAM. The interest among law enforcement to upskill their policing abilities is there, but the tools and the training appear to lag.

Governments

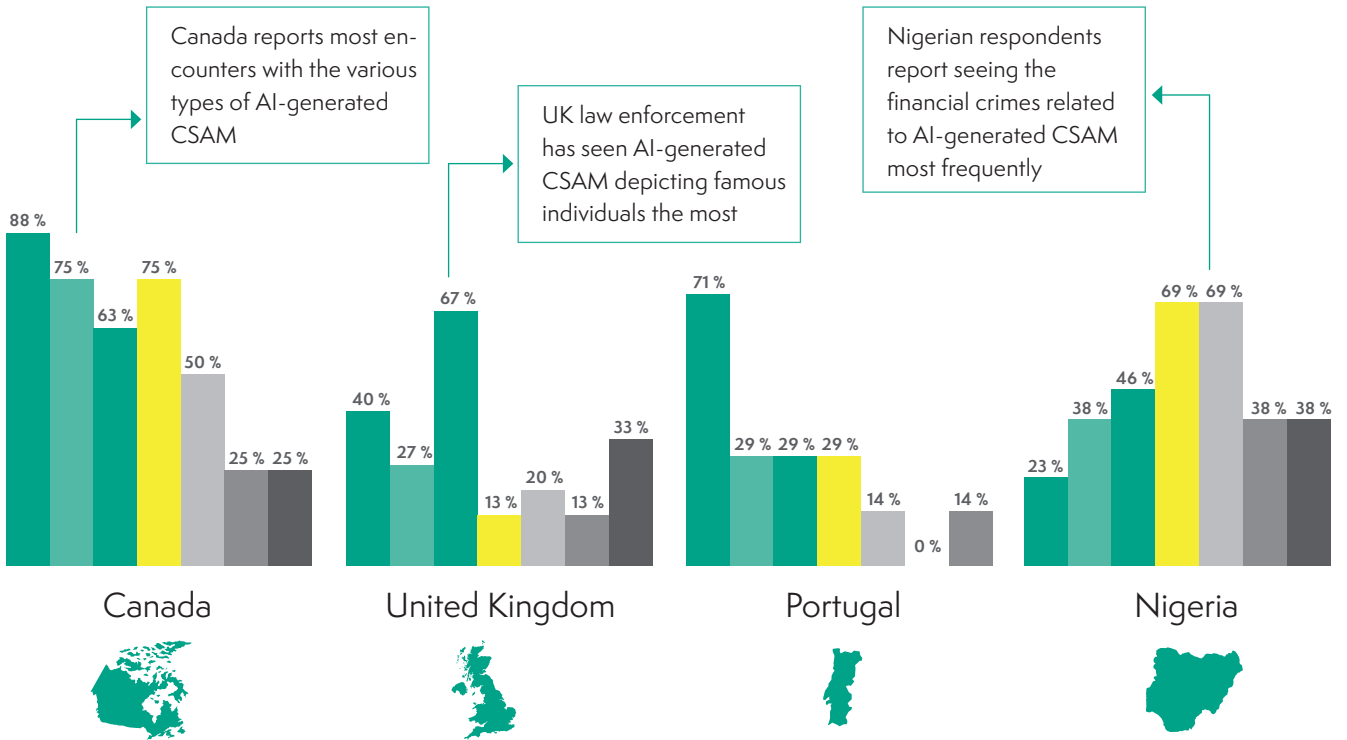
Although some governments have adopted new legislation and signed international declarations pledging their support for countering AI-generated CSAM, the threat appears to remain low on the list of governmental priorities. According to Oldroyd, the motivations behind this may, at least in part, stem from considerations in political communities that governments have more to fear from the use of AI to spread election misinformation, deepfakes of politicians, or radicalizing content than they do from AI-generated CSAM.²¹³

Governments are also contending with how they can future-proof their legislation. Expanding definitions and creating laws specifically tailored to AI-generated CSAM, as discussed in

²¹² Attwood & Bailey, interview.

²¹³ Oldroyd, interview.

FIGURE 10: LAW ENFORCEMENT’S RESPONSE TO WHICH TYPES OF AI-GENERATED CSAM AND RELATED CRIMES THEY HAVE ENCOUNTERED, BY COUNTRY²¹⁴



Question: Which of the following [types of AI-generated CSAM and crimes related to AI-generated CSAM] have you encountered?

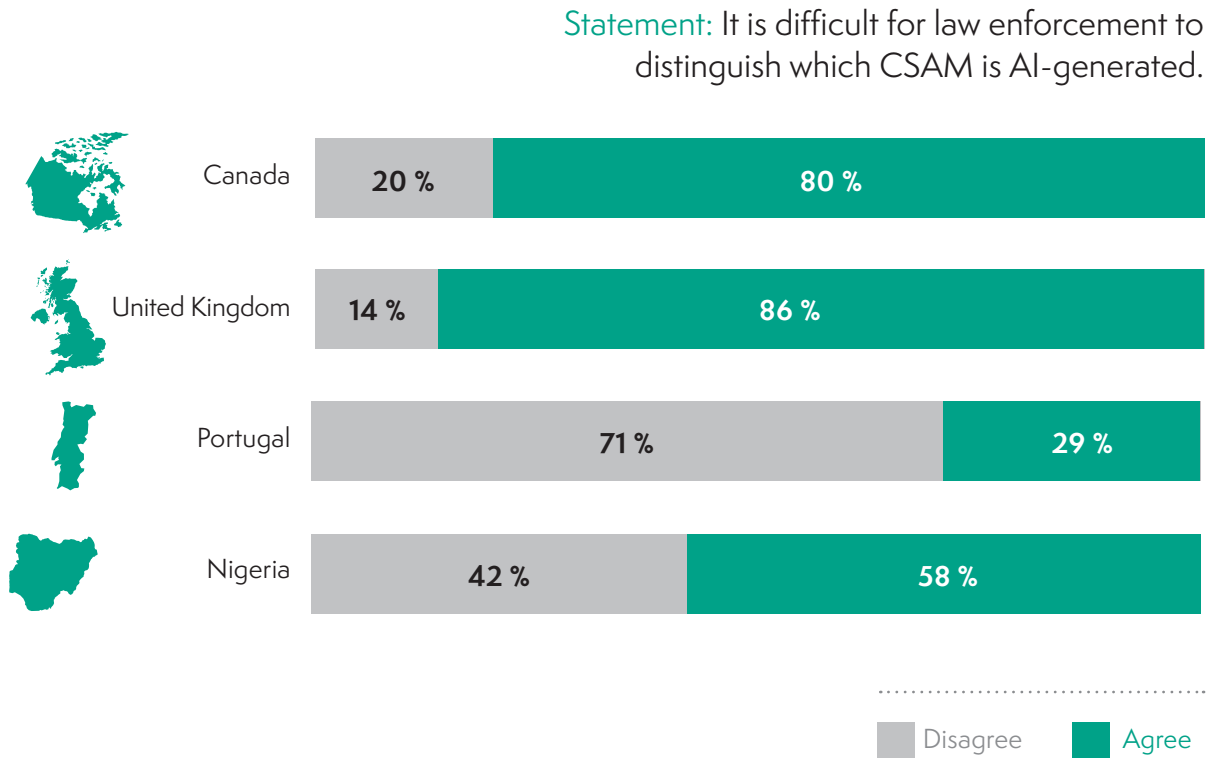
- "nudification" AI-tools used on minors
- AI used to alter or manipulate CSAM of existing victims
- AI-generated CSAM depicting famous individuals
- AI-generated CSAM used for blackmail or extortion
- AI-generated CSAM used for online grooming or enticement of minors
- online platforms refusing to remove AI-generated CSAM
- online platforms selling models/instructions to create AI-generated CSAM

earlier sections, will no doubt protect children in the immediate term. However, truly future-proofed legislation may require an overhaul of the way in which the internet and tech platforms are treated,

particularly in jurisdictions such as the United States, where the size and the tech industry has wide-ranging effects on the rest of the world.

²¹⁴ Value for Good Survey, 2024. Four countries profiled (Canada, United Kingdom, Portugal, Nigeria) were chosen based on highest response rates to survey (10, 15, 9, 17, respectively).

FIGURE 11: LAW ENFORCEMENT’S PERCEIVED ABILITY TO DISTINGUISH AI-GENERATED CSAM BY COUNTRY²¹⁸



The U.S. regulation surrounding NCMEC’s CyberTipline is one example of how the detection of online and digital child abuse crimes may be hindered by systems that were not designed to address them. For instance, even if tech platforms successfully detect CSAM, the U.S. process for reporting a CyberTip can slow down investigations. Due to U.S. constitutional protections, namely the Fourth Amendment right to protection against unreasonable searches and seizures, NCMEC and law enforcement are unable to open CyberTips unless the ESP has indicated that they have already seen the content or law enforcement obtains a warrant, which can take weeks.²¹⁵ In 2024 a report by the Stanford Internet Observatory detailed the limitations of the CyberTipline due to regulations that largely slow down the reporting process.²¹⁶ The REPORT Act, signed into law in April 2024, should ameliorate some of the concerns by instituting

a longer hold period for evidence, but does not completely overhaul the system.²¹⁷

As the headquarters of many of the world’s most-used tech platforms, U.S. law has wide-reaching influence over online safety. The United States government will need to wade into the issue of Section 230 of the 1996 Communications Decency Act and its relationship to AI-generated content. Section 230 is considered foundational to the modern internet, declaring that in the United States, “No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.”²¹⁹ By not holding platforms responsible for what is posted on them by “content creators,” providers like Meta and Google have less of an incentive to monitor the content posted on their sites.

²¹⁵ McQue, “Revealed: US Police Prevented.”

²¹⁶ Grossman et al., “Online Child Safety Ecosystem.”

²¹⁷ Lauren Forristal, “Biden Signs Bill to Protect Children from Online Sexual Abuse and Exploitation,” *Techcrunch*, May 7, 2024.

²¹⁸ Value for Good Survey, 2024.

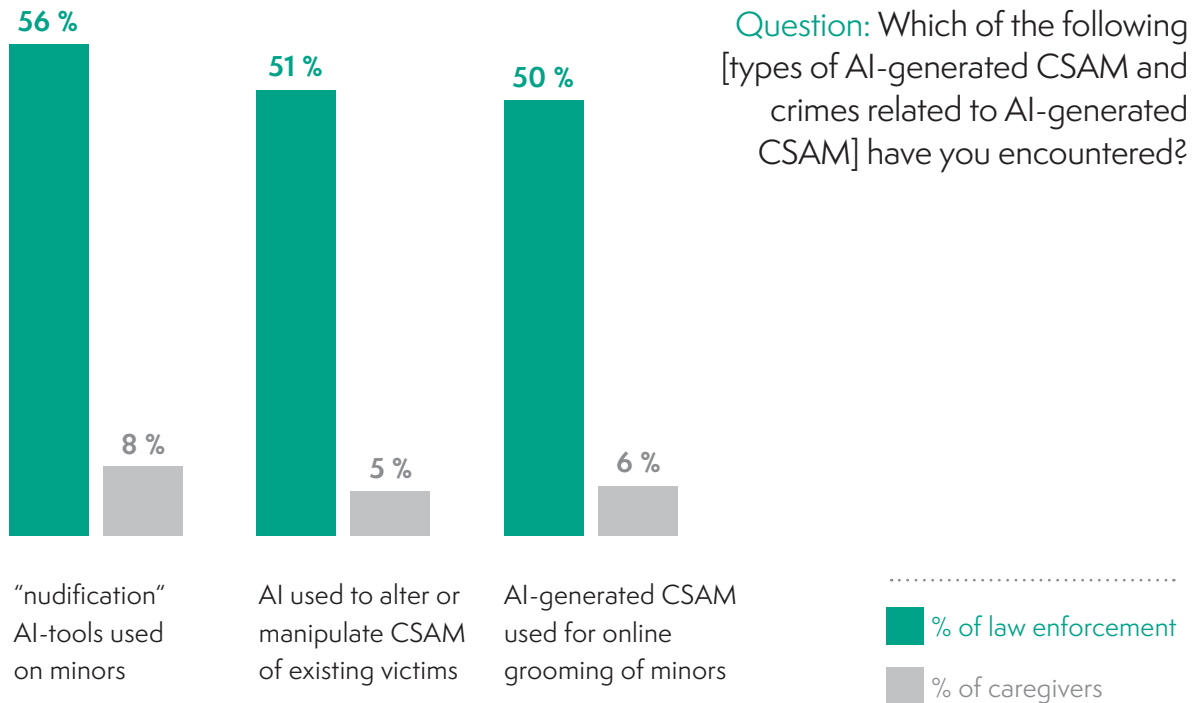
²¹⁹ Protection for Private Blocking and Screening of Offensive Material, U.S. Code 47 (2024) § 230.

In the 1990s the fear was that any broad language requiring internet providers to censor their platforms would constrain the nascent internet.²²⁰ A similar case could arise with AI and its effects on CSAM. If the AI models that can create CSAM are treated as providers and the blame for offending behavior lies with the individuals who write the explicit prompts, AI will facilitate the spread of CSAM, just as the internet has done. The “wait-and-see” approach adopted by many governments shows a lack of urgency in addressing the topic in a systemic manner and could open the door for further victimization.

 Caregivers

While governments, the private sector, and law enforcement grapple with how to slow the production and distribution of AI-generated CSAM, caregivers are faced with the very immediate dilemma of how to protect the children in their care. To understand how caregivers are responding to this threat and what support they need to better protect their children, a survey of 103 parents, teachers, administrators, and medical professionals was conducted in the spring of 2024 for this report. The results of this survey, as summarized below, provide unique insights into how caregivers are contending with online child safety in the age of AI.

FIGURE 12: COMPARISON BETWEEN LAW ENFORCEMENT’S AND CAREGIVERS’ EXPERIENCE WITH AI-GENERATED CSAM²²¹



²²⁰ Alan Rozenshtein, “Interpreting the Ambiguities of Section 230,” Brookings, October 26, 2023.

²²¹ Value for Good Survey, 2024.



Awareness

Caregivers' awareness of the threat of AI-generated CSAM is uneven, as is their experience with AI in general. Nearly 80% of respondents reported being at least somewhat familiar with the concepts of AI and machine learning; however, only 20% of respondents indicated that they are at least somewhat familiar with AI-generated CSAM. This suggests that caregivers are largely not aware of the harms that could befall their children. In fact, 54% of respondents had never heard of "nudify" apps, which are perhaps the most prevalent danger for teenagers in schools. Nearly 70% of respondents were aware of the potential for AI to turn innocuous photos of children sexual, a positive sign given the rise of "sharenting" on social media. AI-generated CSAM is already being seen and discussed in many circles; 7% of the parents who responded to the survey had encountered instances of AI-generated CSAM. This statistic, while low, does indicate that AI-generated CSAM is spreading into the mainstream.

Perhaps unsurprisingly, caregivers report seeing instances of AI-generated CSAM and its effects at significantly lower rates than law enforcement. Still, that over 5% of respondents have come into contact with "nudify" apps, AI used to manipulate existing CSAM, and AI-generated CSAM used for online grooming indicates that this content is bleeding into mainstream consciousness.

Perception of the Challenge

Awareness of the problem is just one part of caregivers' response to the issue of AI-generated CSAM. Their perception of the challenge is another. In the responses to the survey, three themes emerged:

1. A central question caregivers must answer for themselves is **how much unsupervised access** they give their children to internet-connected devices. They must balance the potential dangers that come with internet access with the lack of connectivity with friends and contemporary

culture that results from keeping children away from the internet. Caregivers responding to the survey were concerned that by not allowing children to access the internet they will be stunting their development and depriving them of the privacy to develop independently.

2. Caregivers also reported worrying about their **children's technological savvy** outpacing theirs, enabling their children to get around whatever controls the caregivers set. Caregivers acknowledge that in a rapidly changing technological environment, their children can keep up faster than they can. Moreover, caregivers are concerned that they will be operating with outdated information that does not reflect the dangers facing their children today. They fear that if they do not know how AI-generated CSAM is created or hosted, they will not be able to protect their children. Not understanding the threat or what their children are seeing online gives many caregivers the feeling that the danger is too big for them to tackle.
3. A third challenge caregivers identified in the survey is the **social media platforms** themselves. Caregivers recognize the threat that these platforms pose, from the proliferation of "adult content" to the algorithms that are designed to push, sometimes harmful, images and text to their children's screens. Caregivers are worried about the lack of control they have over what their children are seeing and interacting with and do not feel that the parental controls offered are sufficient.

These responses indicate that caregivers are broadly aware that internet access poses dangers for their children. There is a common understanding of what challenges adults and children face in navigating the digital environment, but a precise understanding of the threat of AI-generated CSAM has not reached mainstream consciousness.



Attitudes Towards the Solution Landscape

Caregivers realize that a primary hurdle to protecting their children is their own “apathy,” “complacency,” and “denial” regarding the situation. Three-quarters of the respondents highlighted the importance of education for parents and other caregivers on how to deal with AI-generated CSAM. As with law enforcement, there is a need for training for caregivers on how to better protect children. Only 16% of respondents reported that they had received training related to child protection or online safety. The majority of the trainings were mandated reporter trainings, which are designed for professionals who are legally required to report maltreatment of a child to relevant authorities. No respondent identified a training that specifically dealt with the negative ramifications of AI or social media.

Caregivers also emphasized educating children on the risks of engaging with people, especially strangers, on the internet. One caregiver stressed the importance of who is delivering that message of caution to children:

“The most impactful training will be directed at children in a manner they are receptive to ([the] messenger is as important as the message). The other most effective counter[measure] will come from the makers of the technology or social media platforms, [which] are allowing the harmful content to spread. I am not very hopeful [about] the latter.”

Teaching children to be cautious about websites they visit and the content they consume will be vital. Whether at home or school, caregivers noted that these conversations need to be had with children.

Survey respondents were more conflicted regarding the efficacy of monitoring their children’s devices as a solution to tackling AI-generated CSAM:

- 35% recommended enabling parental controls on devices and internet browsers to restrict access to age-inappropriate content

- 24% were in favor of using monitoring software to track children’s online activities and receive alerts about potentially harmful content

Monitoring children’s usage of social media and the internet is challenging, as it treats them as the untrustworthy actors, rather than addressing the online predators. However, at a time when the structures are not available to protect children, monitoring their usage of the internet may be necessary to keep them safe.

Although some respondents seemed pessimistic about the solutions available for caregivers to protect their children, others offered ideas that could go a long way towards preventing their children from being exposed to or victims of generative AI capabilities. It is clear from the responses though, that caregivers are playing defense against the spread of AI-generated CSAM and feel it is their responsibility to protect their children and enable them to protect themselves. Without a comprehensive safeguarding framework, caregivers will be left to navigate this emerging threat alone.



V. “The Challenge of the Decade”²²²

RECOMMENDATIONS FOR COUNTERING THE THREAT

The response from governments, law enforcement, the private sector, and caregivers shows that there remains much to be done to keep children from becoming victims of AI-generated CSAM. An ideal response would stop not only AI-generated CSAM and CSAM, but also turn the tide on the normalization of sexualizing children.

At this stage in the development of AI-generated CSAM, the following are recommendations for the various stakeholder groups. As this issue evolves, further recommendations will develop.

Stakeholders & role



Private sector

Inhibiting the creation of AI-generated CSAM largely falls to the private sector. AI developers must implement safety by design measures to prohibit models from generating explicit and violent content, especially of children. This will require high scrutiny of training databases and a variety of other measures that AI developers are best positioned to create and implement. Internet and social media platforms must do their part to cut-off the distribution and commercialization channels of AI-generated CSAM. From blocking websites to content moderation to updating algorithms, tech platforms must put children’s safety before profits.

Responsibilities & recommendations

- Implement safety by design measures to prohibit models from generating explicit or violent content, especially of children
- Ensure high scrutiny of training databases
- Invest in potential solutions, like watermarking AI-generated content
- Update algorithms to stop recommending explicit content & facilitating interaction between potential perpetrators and young children
- Moderate content beyond just flagging images
- Block websites where CSAM is available
- Refuse to advertise on platforms that do not monitor and remove CSAM or AI-generated CSAM



Law enforcement

To effectively respond to the threat of AI-generated CSAM, law enforcement must ensure that they are aware of developments in the criminal landscape through frequent (international) exchange. Moreover, law enforcement must take it upon themselves to explore and take advantage of new tools that are available to them for identifying AI-generated CSAM.

- Ask the question, “has this child ever been abused” when dealing with AI-generated CSAM, thereby expanding the understanding of harm
- Engage in international cooperation: Law enforcement in other countries may be seeing developments in the software before you do
- Stay aware of the developments of AI technologies, both helpful and harmful
- Ensure that the tools that are employed are explainable in courts

Stakeholders & role



Governments

Ensuring that legislation stays up to date with technological advancements is critical to addressing AI-generated CSAM. Policy-makers and regulators must understand the AI-generated CSAM lifecycle to effectively legislate the creation, distribution, and commercialization of this material. Systemic reforms are needed across the board as well as increased investment in technology, training, and education. Moving forward, governments will need to extend collaborations with tech providers to ensure that safeguards for children are put in place.

Responsibilities & recommendations

- Revise legislation to explicitly criminalize AI-generated CSAM
- Hold service providers responsible for hosting content & AI developers for the creation of content
- Consider systemic reforms to how the internet and social media are governed, rather than reactionary legislation to the latest threat
- Engage in partnerships with tech companies, to ensure they are held accountable and are putting safeguards in place
- Invest in advanced technology, education, and training for law enforcement, to ensure perpetrators are not a step ahead technologically
- Invest in solutions for perpetrators who look for help (e.g., propose alternatives to the mandatory reporting that psychiatrists have to incentivize perpetrators to seek help)
- Invest in further education to all stakeholders, from children to adults
- Treat the high offending numbers as a public health issue and address it accordingly



Caregivers

To protect the children in their care, teachers, parents and other caregivers must stay aware of developing threats to children online. They also must discuss the dangers on the internet openly and often, understanding that “a bad conversation is better than no conversation.”²¹³ Moreover, caregivers should actively seek out existing toolkits and advice and consider reducing their child’s online presence.

- Reconsider posting images of your children on social media
- Stay apprised of online dangers for children
- Stay abreast of the latest technology
- Discuss the dangers on the internet openly and often (e.g., speak to children about the consequences of images they post online being altered)
- Make clear to teenagers that “nudify” apps are CSAM
- Talk to children about the dangers of sextortion and know what to do in the event of an incident
- Utilize existing toolkits and advice
- Disseminate information for other caregivers

ACKNOWLEDGMENTS

This report has been produced by **Bracket Foundation** and **Value for Good** in collaboration with the **UNICRI Centre for AI and Robotics**.

The following individuals contributed with interviews and input:

Agustina Callegari

World Economic Forum

Mike Frend

U.K. Online CSEA Covert Intelligence Team

Regina Jensdottir

Council of Europe

Alexander Seger

Council of Europe

Minh Dao

Chainalysis

Sam Gregory

Witness.org

David Haddad

Phoenix Police Department

Natalia Bayona

United Nations Tourism

Simon Bailey

Child Rescue Coalition

Marco Fontani

Amped Software

Nate King

International Justice Mission

Tom Oldroyd

Semantics21

Martino Jerian

Amped Software

Phil Attwood

Child Rescue Coalition

Zara Gasparyan

Council of Europe



ABOUT THE AUTHORS

Bracket Foundation is the philanthropic venture arm of Bracket Capital, a private markets investor based in Los Angeles with offices in Doha and London. Bracket Foundation's mission is to harness the power of technology for social good by leveraging technology solutions to tackle growing global challenges. In 2019, alongside its longstanding partners, Bracket Foundation issued a leading publication on how Artificial Intelligence can combat online sexual child abuse. The publication was presented on the sidelines of the United Nations General Assembly that same year and resulted in a multi-year partnership with UNICRI (the United Nations Interregional Crime and Justice Research Institute). The partnership served as the backbone for the UN sponsored "AI for Safer Children" platform that ensued to empower law enforcement agencies worldwide with innovative tools to better detect, prevent and prosecute online sexual abuse being committed against children. In 2022, Bracket Foundation published the sequel to its initial whitepaper which explores the dangers and vulnerabilities on children of the Metaverse and Online Social Gaming Platforms. The paper prompted Big Tech companies to review their user experience practices to make sure their products were safer by design for children and teens. Bracket Foundation is engaged with several public sector actors which include multilateral organizations (such as the UN, the European Union, the European Commission), NGOs and States to raise awareness on the uses of AI, building trust between the public and private sector, promoting more government investment in AI, and lobbying for changes to the legislative framework around data use in order to scale a global solution to online child safety. Bracket Foundation is also engaged in advocacy work especially as it relates to holding Big Tech companies accountable for safer platforms.

Yalda Aoukar is the Co-Founder and Managing Partner of Bracket Capital, a private markets investor based in Los Angeles, CA established in 2016. Bracket Capital pursues a value driven strategy in growth and late-stage technology companies optimizing for asymmetries and dislocations that exist in private markets. Yalda is the President of Bracket Foundation, the philanthropic arm of Bracket Capital with the mission to leverage the power of technology for social good, especially as it relates to solving the world's most pressing global challenges. Since inception, Bracket Foundation has focused its efforts heavily on online child safety having cemented its role as an early thought-leader in the space. Yalda is a champion for women's empowerment and entrepreneurship in Venture Capital and other financial sectors, where women are traditionally underrepresented. In addition to investing in leading technology companies, she serves as an adviser to governments and policy makers on digital development in diverse fields such as Biotech, Food Security, Artificial Intelligence Integration and Education Technology. She sits on the board of the United Nation's AI for Safer Children Initiative which she helped launch in 2020, as well as the World Innovation Summit for Education (WISE) Accelerator. Yalda is a Young Global Leader of the World Economic Forum (Class of 2024). Yalda holds a Master in Public Policy (MPP) from the Harvard Kennedy School of Government (Class of 2008). She spends her time between London, Los Angeles and the Middle East.

UNICRI Centre for AI & Robotics is a specialized center of the United Nations Interregional Crime and Justice Research Institute (UNICRI). It was established in 2017 and is located in The Hague, the Netherlands. Its mission is to advance understanding of the challenges and opportunities brought by AI and other new and emerging technologies from the perspective of justice, crime and other security threats. It seeks to support Member States better to understand the benefits and risks of these technologies and leverage their potential in a responsible manner.



Maria Eira is the Artificial Intelligence Expert at UNICRI's Centre for AI and Robotics. At the Centre, she provides technical advice to ongoing projects on the benefits and risks of AI and emerging technologies. In particular, her work focuses on the issues of crime prevention, criminal justice, and the rule of law.

Value for Good GmbH is a consultancy specialized in the field of social impact that envisions a world in which effective action is taken to solve societal challenges. To achieve this Value for Good inspires through insights, advises through consulting and empowers through training. Value for Good serves leaders from private sector, governments, international institutions, foundations and nonprofits and equips them with the knowledge and tools to make a positive and sustainable difference.

Clara Péron is the founder and managing director of Value for Good. Originally from Montréal, Canada, she has lived and worked internationally—

with longer postings in India, Cambodia, Egypt, Ukraine, the United States, and Germany. Clara started her career in 2002 in the Canadian foreign service and after working as a strategy consultant for the Boston Consulting Group's Berlin office she founded Value for Good and the Value for Good Foundation.

Lisa Maddox is a principal at Value for Good with a focus on Tech for Good. In addition, she is the co-lead of the Global Development practice group at Value for Good. Prior to joining Value for Good, she was the Chief of Staff at Fuzu, a Nairobi-based job-tech start up, and a project leader at Accenture in New York.

Lana Apple is a consultant at Value for Good with a background in education research, education policy, and the rights of the child. She holds a master's degree in international and comparative education from the University of Oxford. Lana has taught in the United States and Germany and published in the field of international education.

FURTHER READING

Grossman, Shelby, et al. "The Strengths and Weaknesses of the Online Child Safety Ecosystem." Stanford Internet Observatory. 2024. <https://stacks.stanford.edu/file/druid:pr592kc5483/cybertipline-paper-2024-04-22.pdf>.

International Justice Mission and University of Nottingham Rights Lab. "Scale of Harm Research Method, Findings, and Recommendations: Estimating the Prevalence of Trafficking to Produce Child Sexual Exploitation Material in the Philippines." 2023. https://assets.ijm.app/IJM_Scale_of_Harm_2023_Full_Report_5f292593a9.pdf.

Internet Watch Foundation. "How AI is Being Abused to Create Child Sexual Abuse Imagery." IWF. October 2023. https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf.

Latakos, Santiago. "A Revealing Picture." Graphika. 2023. <https://public-assets.graphika.com/reports/graphika-report-a-revealing-picture.pdf>.

Thiel, David. "Identifying and Eliminating CSAM in Generative ML Training Data and Models." Stanford Internet Observatory, 2023. https://stacks.stanford.edu/file/druid:kh752sm9123/ml_training_data_csam_report-2023-12-23.pdf.

Thorn and NCMEC. "Trends in Financial Sextortion. An Investigation of Sextortion Reports in NCMEC CyberTipline Data." 2024. https://info.thorn.org/hubfs/Research/Thorn_TrendsInFinancialSextortion_June2024.pdf.

BIBLIOGRAPHY

- AIAAIC. "AIAAIC Repository." Accessed April 11, 2024. <https://www.aiaaic.org/>.
- BBC. "Japan Bans Child Pornography Possession." *BBC*. June 18, 2014. <https://www.bbc.com/news/world-asia-27898841>.
- Beaumont, Romain. "LAION-5B: A New Era of Open Large-Scale Multimodal Datasets." LAION.ai. March 2022. <https://laion.ai/blog/laion-5b/>.
- Berman, Marc. "Bill to Protect Children from AI Enabled Sexual Exploitation Passes Assembly." Press Release. May 23, 2024. [https://a23.asmdc.org/press-releases/20240523-bill-protect-children-ai-enabled-sexual-exploitation-passes-assembly#:~:text=Bill%20to%20Protect%20Children%20from%20AI%20Enabled%20Sexual%20Exploitation%20Passes%20Assembly,-For%20immediate%20release&text=SACRAMENTO%20%E2%80%93%20Today%20the%20Assembly%20passed,Sexual%20Abuse%20Material%20\(CSAM\)](https://a23.asmdc.org/press-releases/20240523-bill-protect-children-ai-enabled-sexual-exploitation-passes-assembly#:~:text=Bill%20to%20Protect%20Children%20from%20AI%20Enabled%20Sexual%20Exploitation%20Passes%20Assembly,-For%20immediate%20release&text=SACRAMENTO%20%E2%80%93%20Today%20the%20Assembly%20passed,Sexual%20Abuse%20Material%20(CSAM)).
- Blunt, Katherine & Horwitz, Jeff. "Instagram Connects Vast Pedophile Network." *The Wall Street Journal*. June 7, 2023. <https://www.wsj.com/articles/instagram-vast-pedophile-network-4ab7189>.
- Bracket Foundation. "Artificial Intelligence: Combating Online Sexual Abuse of Children. Bracket Foundation." 2019. <https://static1.squarespace.com/static/5d7cd3b6974889646fce45c1/t/632f37b896470d1340136fc9/1664038845748/AI.pdf>.
- Bracket Foundation. "Gaming and the Metaverse: The Alarming Rise of Online Sexual Exploitation and Abuse of Children Within the New Digital Frontier." Bracket Foundation, 2022. <https://static1.squarespace.com/static/5d7cd3b6974889646fce45c1/t/632f3344e-acdbb108c8c356f/1664037701806/metaverse+%26+gaming.pdf>.
- Briggs, Peter, Walter T. Simon, and Stacy Simonsen. "An Exploratory Study of Internet-Initiated Sexual Offenses and the Chat Room Sex Offender: Has the Internet Enabled a New Typology of Sex Offender?" *Sexual Abuse: A Journal of Research and Treatment*. No. 1(2011): 23. P. 72 -91. <https://www.dhi.ac.uk/san/waysofbeing/data/communication-zangana-briggs-2011.pdf>.
- Butler, Josh. "Search Engines Required to Stamp out AI-Generated Images of Child Abuse Under Australia's New Code." *The Guardian*. September 7, 2023. <https://www.theguardian.com/technology/2023/sep/08/search-engines-required-to-stamp-out-ai-generated-images-of-child-abuse-under-australias-new-code>.
- Chainalysis. *The 2024 Crypto Crime Report*. 2024. <https://www.chainalysis.com/wp-content/uploads/2024/06/the-2024-crypto-crime-report-release.pdf>.
- Chriyst, Tommy. "Reflecting on the 2024 AI Safety Summit and What It Means." LinkedIn. May 22, 2024. <https://www.linkedin.com/pulse/reflecting-2024-ai-safety-summit-what-means-tommy-chriyst-wvk6e/>.
- Das, Shanti. "Sex Offender Banned from Using AI Tools in Landmark UK Case." *The Guardian*. April 21, 2024. <https://www.theguardian.com/technology/2024/apr/21/sex-offender-banned-from-using-ai-tools-in-landmark-uk-case>.
- Data Commons. "World Demographics." Datacommons.org. Accessed June 2, 2024. <https://datacommons.org/explore/#q=country%20populations>.
- Dwoskin, Elizabeth & Gerrit De Vynck. "Facebook's Internal Chat Boards Show Politics Often at Center of Decision Making." *The Washington Post*. October 24, 2021. <https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/>.
- ECPAT International. "Luxembourg Guidelines: Terminology Guidelines for the Protection of Children from Sexual Exploitation and Sexual Abuse." Interagency Working Group on Sexual Exploitation of Children. January 28, 2016. <https://ecpat.org/wp-content/uploads/2021/05/Terminology-guidelines-396922-EN-1.pdf>.
- ECPAT International, "Intervention on Cybercrimes Against Children, Including CSAM," United Nations Office on Drugs and Crime, N.d. https://www.unodc.org/documents/Cybercrime/AdHocCommittee/Third_intersessional_consultation/Statements-submissions/Cybercrimes_ECPAT_CSAM_intervention_final.pdf.
- Edwards, Benj. "New Stable Diffusion 3 Release Excels at AI-Generated Body Horror." *ARS Technica*. June 12, 2024. <https://arstechnica.com/information-technology/2024/06/ridiculed-stable-diffusion-3-release-excels-at-ai-generated-body-horror/>.
- European Commission. "Q&A - The Fight Against Child Sexual Abuse Receives New Impetus with Updated Criminal Law Rules." Press Corner. February 6, 2024. https://ec.europa.eu/commission/presscorner/detail/en/qanda_24_643.
- Federal Bureau of Investigation. Department of Justice. "Child Sexual Abuse Material Created by Generative AI and Similar Online Tools is Illegal." March 29, 2024. <https://www.ic3.gov/Media/Y2024/PSA240329#fn1>.
- Fitzgerald, Madyson. "States Race to Restrict Deepfake Porn as it Becomes Easier to Create." *Utah News Dispatch*. April 12, 2024. <https://utahnewsdispatch.com/2024/04/12/states-race-to-restrict-deepfake-porn-as-it-becomes-easier-to-create/#:~:text=%E2%80%9CWhat%20we%20should%20be%20focusing,who%20are%20making%20the%20software.%E2%80%9D>.

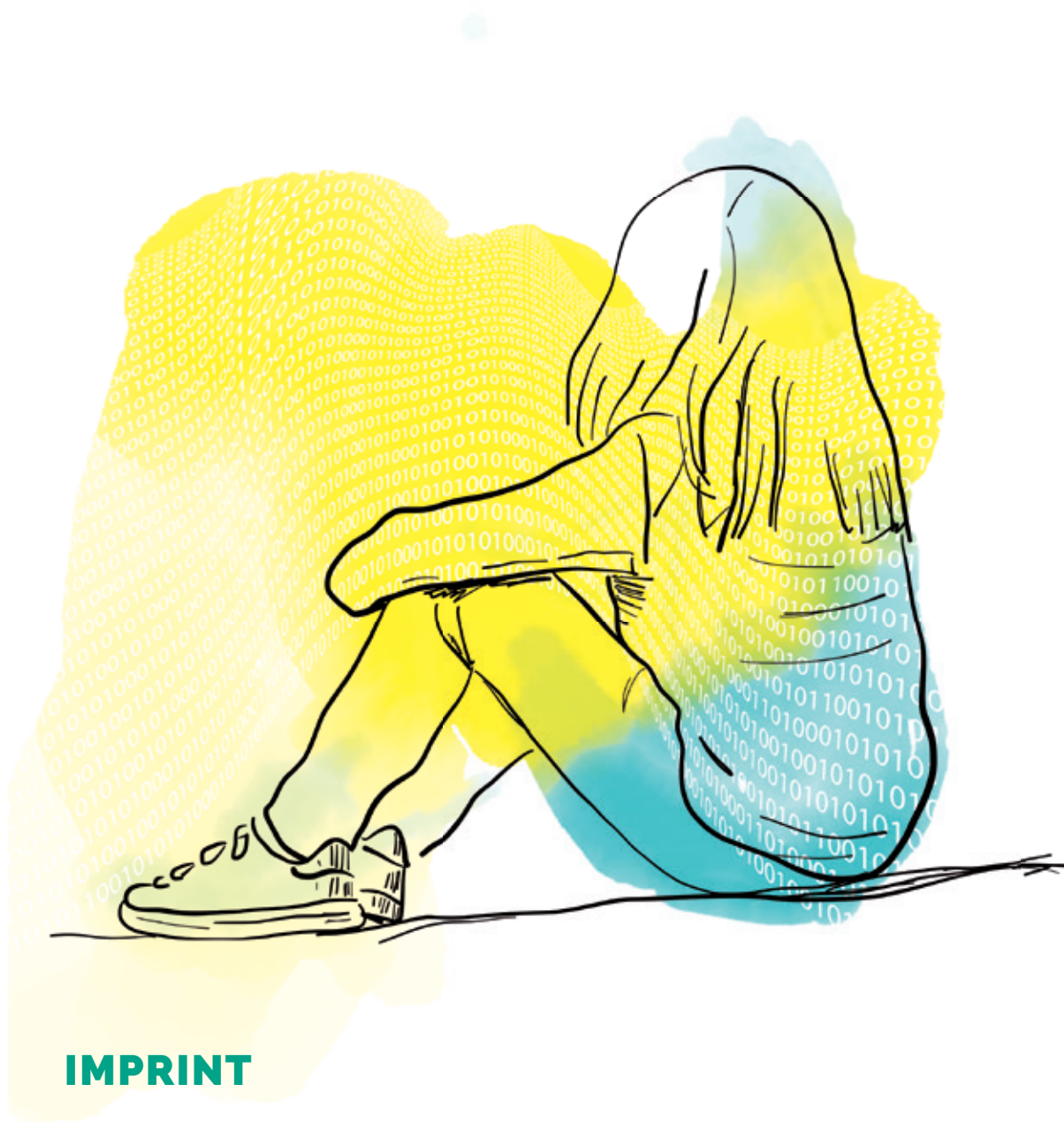
- Forristal, Lauren. "Biden Signs Bill to Protect Children from Online Sexual Abuse and Exploitation." *Techcrunch*. May 7, 2024. <https://techcrunch.com/2024/05/07/biden-signs-bill-to-protect-children-from-online-sexual-abuse-and-exploitation/?guccounter=1>.
- GIP Digital Watch Observatory. "Hiroshima AI Process: G7's Effort to Tackle Challenges of AI Technology." Dig.Watch. May 21, 2023. <https://dig.watch/updates/hiroshima-ai-process-g7s-effort-to-tackle-challenges-of-ai-technology>.
- Government of Canada. "Proposed Bill to Address Online Harms." Accessed July 8, 2024. <https://www.canada.ca/en/canadian-heritage/services/online-harms.html>.
- Government of the United Kingdom. "A Joint Statement from the United States and the United Kingdom on Combatting Child Sexual Abuse and Exploitation." Gov.uk. September 27, 2023. <https://www.gov.uk/government/publications/combating-child-sexual-abuse-and-exploitation/a-joint-statement-from-the-united-states-and-the-united-kingdom-on-combating-child-sexual-abuse-and-exploitation>.
- Government of the United Kingdom. "Countries Agree to Safe and Responsible Development of Frontier AI in Landmark Bletchley Declaration." Gov.uk. November 2023. <https://www.gov.uk/government/news/countries-agree-to-safe-and-responsible-development-of-frontier-ai-in-landmark-bletchley-declaration>.
- Government of the United Kingdom. "Government Cracks Down on 'Deepfakes' Creation." Gov.uk, Ministry of Justice. April 16, 2024. https://www.gov.uk/government/news/government-cracks-down-on-deepfakes-creation?utm_source=ActiveCampaign&utm_medium=email&utm_content=Global%20legislation%20and%20policy%20changes%20keeping%20children%20safe%20online&utm_campaign=Legislative%20%26%20Policy%20Update%20June%202024.
- Government of the United Kingdom. "Home Secretary Urges Meta to Protect Children from Sexual Abuse." Gov.uk. September 20, 2023. <https://www.gov.uk/government/news/home-secretary-urges-meta-to-protect-children-from-sexual-abuse>.
- Government of the United Kingdom. "UK and US Pledge to Combat AI-Generated Images of Child Abuse." Gov.uk. September 27, 2023. <https://www.gov.uk/government/news/uk-and-us-pledge-to-combat-ai-generated-images-of-child-abuse>.
- Gregory, Andy. "Prioritise Children's Online Safety at Election to Tackle 'Hidden Pandemic' of Sexual Abuse, Experts Urge." *The Independent*. June 2, 2024. <https://www.independent.co.uk/news/uk/politics/sexual-abuse-children-online-safety-childlight-b2552670.html>.
- Grossman, Shelby, Riana Pfefferkorn, David Thiel, Sara Shah, Alex Stamos, Renée DiResta, John Perrino, Elena Cryst, & Jeffrey Hancock. "The Strengths and Weaknesses of the Online Child Safety Ecosystem." Stanford Internet Observatory. April 22, 2024. <https://stacks.stanford.edu/file/druid:pr592kc5483/cybertipline-paper-2024-04-22.pdf>.
- Harwell, Drew & Pranshu Verma. "In Novel Case, U.S. Charges Man with Making Child Sex Abuse Images with AI." *The Washington Post*. May 21, 2024. <https://www.msn.com/en-us/news/us/in-novel-case-us-charges-man-with-making-child-sex-abuse-images-with-ai/ar-BB1mO208>.
- Haskins, Caroline. "Florida Middle Schoolers Arrested for Allegedly Creating Deepfake Nudes of Classmates." *Wired*. March 8, 2024. <https://www.wired.com/story/florida-teens-arrested-deepfake-nudes-classmates/>.
- Hedgecoe, Guy. "AI-Generated Naked Child Images Shock Spanish Town of Almendralejo." *BBC*. September 24, 2023. <https://www.bbc.com/news/world-europe-66877718>.
- Hemrajani, Asha. "China's New Legislation on Deepfakes: Should the Rest of Asia Follow Suit?" *The Diplomat*. March 8, 2023. <https://thediplomat.com/2023/03/chinas-new-legislation-on-deepfakes-should-the-rest-of-asia-follow-suit/>.
- Hern, Alex. "Can AI Image Generators be Policed to Prevent Explicit Deepfakes of Children?" *The Guardian*. April 23, 2024. <https://amp.theguardian.com/technology/2024/apr/23/can-ai-image-generators-be-policed-to-prevent-explicit-deepfakes-of-children>.
- Home Security Heroes. "2023 State of Deepfakes. Realities, Threats, and Impact." 2023. <https://www.homesecurityheroes.com/state-of-deepfakes/#key-findings>.
- Hudson, Valerie. "The Right Way to Deal with AI-Generated Child Pornography." *Deseret News*. June 2, 2024. <https://www.deseret.com/opinion/2024/06/02/ai-child-sexual-abuse-material-law-germany/>.
- Human Rights Watch. "Brazil: Children's Personal Photos Misused to Power AI Tools." HRW. June 10, 2024. <https://www.hrw.org/news/2024/06/10/brazil-childrens-personal-photos-misused-power-ai-tools>.
- Insoll, Tegan, Anna Ovaska & Nina Vaaranen-Valkonen. "CSAM Users in the Dark Web." Suojellaan Lapsia / Protect Children. 2021. <https://www.suojellaanlapsia.fi/en/post/csam-users-in-the-dark-web-protecting-children-through-prevention>.
- International Centre for Missing & Exploited Children. "Child Sexual Abuse Material: Model Legislation & Global Review." ICMEC, 2023. Tenth edition. https://cdn.icmec.org/wp-content/uploads/2023/10/CSAM-Model-Legislation_10th-Ed-Oct-2023.pdf.
- International Labour Organization. "Profits and Poverty: The Economics of Forced Labour." ILO. March 19, 2024. <https://www.ilo.org/publications/major-publications/profits-and-poverty-economics-forced-labour>.

- Internet Watch Foundation. "How AI is Being Abused to Create Child Sexual Abuse Imagery." IWF, 2023. <https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report-public-oct23v1.pdf>.
- Internet Watch Foundation. "Protect your Customers, Staff & Platform from Websites Showing the Graphic Sexual Abuse of Children with our URL List." IWF, 2024. <https://www.iwf.org.uk/our-technology/our-services/url-list/>.
- Internet Watch Foundation. "Total Number of CSAM Reports." IWF Annual Report, 2021. <https://annualreport2021.iwf.org.uk/trends/total>.
- "Into the Light Index." Childlight Global Child Safety Institute. Accessed June 12, 2024. <https://intothelight.childlight.org/>.
- Jargon, Julie. "Fake Nudes of Real Students Cause an Uproar at a New Jersey High School." *The Wall Street Journal*. November 2, 2023. <https://www.wsj.com/tech/fake-nudes-of-real-students-cause-an-uproar-at-a-new-jersey-high-school-df10f1bb>.
- Jasper, Susan. "An Update on our Child Safety Efforts and Commitments." Google Blog, Safety & Security. April 23, 2024. <https://blog.google/technology/safety-security/an-update-on-our-child-safety-efforts-and-commitments/>.
- Jones, Sam. "Spain Sentences 15 Schoolchildren Over AI-generated Naked Images." *The Guardian*. July 9, 2024. <https://www.theguardian.com/world/article/2024/jul/09/spain-sentences-15-school-children-over-ai-generated-naked-images>.
- Kang, Cecilia & Adam Satariano. "As A.I. Booms, Lawmakers Struggle to Understand the Technology." *The New York Times*. March 3, 2023. <https://www.nytimes.com/2023/03/03/technology/artificial-intelligence-regulation-congress.html>.
- Latakos, Santiago. "A Revealing Picture." *Graphika*. December 2023. <https://public-assets.graphika.com/reports/graphika-report-a-revealing-picture.pdf>.
- Lavinia, Emilie. "'I've Seen Boys Request Fake Nudes of Their Teachers and Mothers': How Nudify Apps Are Violating Women and Girls in the UK." *Glamour*. June 24, 2024. <https://www.glamourmagazine.co.uk/article/nudify-apps-investigation>.
- Levine, Alexandra. "'I Want That Sweet Baby': AI-Generated Kids Draw Predators on TikTok and Instagram." *Forbes*. May 20, 2024. <https://www.forbes.com/sites/alexandralevine/2024/05/20/ai-generated-kids-tiktok-instagram-social-mediachild-safety-predators/>.
- Levine, Alexandra. "Stable Diffusion 1.5 Was Trained on Illegal Child Sexual Abuse Material, Stanford Study Says." *Forbes*. March 25, 2024. <https://www.forbes.com/sites/alexandralevine/2023/12/20/stable-diffusion-child-sexual-abuse-material-stanford-internet-observatory/>.
- Linebaugh, Kate, host. "Teens Are Falling Victim to AI Fake Nudes". WSJ Podcasts. July 12, 2024. <https://www.wsj.com/podcasts/the-journal/teens-are-falling-victim-to-ai-fake-nudes/8f753386-764d-431d-a2de-bc382c89f832>.
- Lomas, Natascha. "Google's Call-Scanning AI Could Dial up Censorship by Default, Privacy Experts Warn." Yahoo! Lifestyle. May 15, 2024. <https://au.lifestyle.yahoo.com/googles-call-scanning-ai-could-172510066.html?guccounter=1>.
- Macchi, Victoria. "Japan Outlaws Owning Child Pornography." Learning English. July 15, 2014. <https://learningenglish.voanews.com/a/japan-outlaw-child-pornography/1955140.html>.
- Martini Grimaldi, Christian. "Pedophilia is not Taboo Enough in Japan." *UCA News*. June 3, 2024. <https://www.ucanews.com/news/pedophilia-is-not-taboo-enough-in-japan/105290>.
- McQue, Katie. "Child Predators Are Using AI to Create Sexual Images of Their Favorite 'Stars': 'My Body Will Never be Mine Again.'" *The Guardian*. June 12, 2024. <https://www.theguardian.com/technology/article/2024/jun/12/predators-using-ai-generate-child-sexual-images>.
- McQue, Katie. "Revealed: US Police Prevented from Viewing Many Online Child Sexual Abuse Reports, Lawyers Say." *The Guardian*. January 17, 2024. <https://www.theguardian.com/technology/2024/jan/17/child-sexual-abuse-ai-moderator-police-meta-alphabet>.
- Metz, Cade & Tiffany Hsu. "OpenAI Releases 'Deepfake' Detector to Disinformation Researchers." *The New York Times*. May 2024. <https://www.nytimes.com/2024/05/07/technology/openai-deepfake-detector.html>.
- Milmo, Dan. "AI Advances Could Lead to More Child Sexual Abuse Videos, Watchdog Warns." July 22, 2024. <https://amp.theguardian.com/society/article/2024/jul/22/ai-child-sexual-abuse-videos-iwf-watchdog>.
- Murphy, Margi. "'Nudify' Apps that Use AI to Undress Women in Photos Are Soaring in Popularity, Prompting Worries about Non-Consensual Porn." *Fortune*. December 8, 2023. https://fortune.com/2023/12/08/ai-pps-undress-women-photos-soaring-in-use/?queryly=related_article.
- National Center for Missing & Exploited Children. *CyberTipline 2023 Report*. 2023. <https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata>.
- National Crime Agency. "NCA Leads International Coalition Tackling Child Sexual Abuse." 2023. <https://nationalcrimeagency.gov.uk/news/nca-leads-international-coalition-tackling-child-sexual-abuse>.
- Nelson, Jason. "Twitter Touts 'Seamless' Blocking of Child Abuse Content as Elon Musk Faces Increased EU Scrutiny." *Emerge*. May 9, 2024. <https://decrypt.co/230013/twitter-elon-musk-child-sexual-abuse-material-csam-elon-musk-eu>.
- Ocasio-Cortez, Alexandria. "Rep. Ocasio-Cortez Leads Bipartisan, Bicameral Introduction of DEFIANCE Act to Combat Use of Non-Consensual, Sexually-Explicit 'Deepfake' Media." Press Release. March 7, 2024. <https://ocasio-cortez.house.gov/media/press-releases/rep-ocasio-cortez-leads-bipartisan-bicameral-introduction-defiance-act-combat>.

- OECD.AI. "OECD AI Incidents Monitor." OECD.AI Policy Observatory. Accessed March 30, 2024 & April 11, 2024. <https://oecd.ai/en/incidents-methodology>.
- Office for National Statistics. "Population Estimates for the UK, England, Wales, Scotland, and Northern Ireland: Mid-2022." ONS. March 26, 2024. <https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2022>.
- Office of Public Affairs, U.S. Department of Justice. "Recidivist Sex Offender Sentenced for Possessing Deepfake Child Sexual Abuse Material." Press Release. May 20, 2024. <https://www.justice.gov/opa/pr/recidivist-sex-offender-sentenced-possessing-deepfake-child-sexual-abuse-material>.
- Office of Public Affairs, U.S. Department of Justice. "Man Arrested for Producing, Distributing, and Possessing AI-Generated Images of Minors Engaged in Sexually Explicit Conduct." Press Release. May 1, 2024. <https://www.justice.gov/opa/pr/man-arrested-producing-distributing-and-possessing-ai-generated-images-minors-engaged>.
- Thorn and NCMEC. "Trends in Financial Sextortion. An Investigation of Sextortion Reports in NCMEC CyberTipline Data." 2024. https://info.thorn.org/hubfs/Research/Thorn_TrendsInFinancialSextortion_June2024.pdf.
- Oremus, Will. "AI Is About to Make the Online Child Sex Abuse Problem Much Worse." *The Boston Globe*. April 22, 2024. <https://www.bostonglobe.com/2024/04/22/nation/ai-is-about-make-online-child-sex-abuse-problem-much-worse/>.
- Popken, Alex. "We're Unprepared for the Threat GenAI on Instagram, Facebook, and WhatsApp Poses to Kids." *Fast Company*. June 7, 2024. <https://www.fastcompany.com/91136311/were-unprepared-for-the-threat-genai-on-instagram-facebook-and-whatsapp-poses-to-kids>.
- Raffile, Paul, Alex Goldenberg, Cole McCann & Joel Finkelstein. "Digital Pandemic: Uncovering the Role of 'Yahoo Boys' in the Surge of Social Media-Enabled Financial Sextortion Targeting Minors." Network Contagion Research Institute, 2024. https://networkcontagion.us/wp-content/uploads/Yahoo-Boys_1.2.24.pdf.
- Roopal. "AI-Generated Child Sexual Abuse: The Threat of Undress AI Tools." *AI Mojo*. June 2024. <https://aimojo.pro/ai-generated-child-sexual-abuse/>.
- Roper, Caitlin. "'No way of Knowing if the Child is Fictional': How 'Virtual' Child Sexual Abuse Material Harms Children." *Collective Shout*. May 6, 2024. https://www.collectiveshout.org/how_virtual_child_sexual_abuse_material_harms_children.
- Rozenshtein, Alan. "Interpreting the Ambiguities of Section 230." *Brookings*. October 26, 2023. <https://www.brookings.edu/articles/interpreting-the-ambiguities-of-section-230/>.
- Russon, Mary-Ann. "Japan: Artist Guilty of Creating Computer-Generated VR Child Pornography of an Actual Girl." *International Business Times*. March 18, 2016. <https://www.ibtimes.co.uk/japan-artist-guilty-creating-computer-generated-vr-child-pornography-actual-girl-1550381>.
- Ryall, Julian. "'Realistic' AI-Generated Child Porn in Japan Sparks Debate on Legal Loophole and 'Kawaii' Culture." *Everand*. November 17, 2023. <https://www.everand.com/article/685275061/Realistic-Ai-Generated-Child-Porn-In-Japan-Sparks-Debate-On-Legal-Loophole-And-Kawaii-Culture>.
- Sandle, Paul. "UK's Online Safety Bill Finally Passed by Parliament." *Reuters*. September 19, 2023. <https://www.reuters.com/world/uk/uks-online-safety-bill-passed-by-parliament-2023-09-19/>.
- Sanzgiri, Vallari. "Here's What the Wisconsin CSAM Incident Tells Us About Prompt-Blocking and E2EE." *Medianama*. June 11, 2024. <https://www.medianama.com/2024/06/223-heres-what-the-wisconsin-csam-incident-tells-us-about-prompt-blocking-and-e2ee/>.
- Saxena, Vishakha. "Japan Leaders Want Law on Generative AI 'Within the Year'." *Asia Financial*. February 15, 2024. <https://www.asiafinancial.com/japan-leaders-want-law-on-generative-ai-within-the-year>.
- Serebrin, Jacob. "Quebec Man who Created Synthetic, AI-Generated Child Pornography Sentenced to Prison." *The Canadian Press*. April 26, 2023. <https://www.cbc.ca/news/canada/montreal/ai-child-abuse-images-1.6823808>.
- Shehan, John. "Addressing Real Harm Done by Deepfakes." Testimony for United States House Committee on Oversight and Accountability. Accessed April 20, 2024. <https://oversight.house.gov/hearing/addressing-real-harm-done-by-deepfakes/>.
- Shimbun, Yomiuri. "Japan Lags in Regulating AI-Generated Child Porn as Loophole in Existing Law Gets Exploited." *The Japan News*. November 14, 2023. <https://japan-news.yomiuri.co.jp/society/general-news/20231114-149299/#:~:text=The%20premise%20of%20the%20law,that%20resemble%20an%20existing%20child>.
- Shinyshiny. "Do AI Images Count as Indecent Images?" *Shinyshiny.tv*. May 15, 2024. <https://www.shinyshiny.tv/2024/05/do-ai-images-count-as-indecent-images.html>.
- Singer, Natascha. "Teen Girls Confront Epidemic of Deepfake Nudes at School." *The New York Times*. April 8, 2024. <https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html>.
- Singh, Sourabh. "Instagram Identifying Photographs that have been Modified with AI." *United Business Journal*. June 20, 2024. <https://theubj.com/technology/instagram-identifying-photographs-that-have-been-modified-with-ai/>.

- Souza, Murilo. "Projeto Prevê Até 8 Anos de Prisão para quem Usar Inteligência Artificial para Gerar Conteúdo Sexual com Crianças." *Camara dos Deputados*. February 27, 2024. <https://www.camara.leg.br/noticias/1038114-projeto-preve-ate-8-anos-de-prisao-para-quem-usar-inteligencia-artificial-para-gerar-conteudo-sexual-com-criancas/>.
- Stability AI. "Stable Diffusion Launch Announcement." *Stability.ai*. August 10, 2023. Accessed August 7, 2024. <https://stability.ai/news/stable-diffusion-announcement>.
- Stanford University Human-Centered Artificial Intelligence. "Artificial Intelligence Index Report 2024." *Stanford Internet Observatory*. 2024. https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf.
- Sudharsan, Deepthi. "Decoding the Relationship: Language Models and Natural Language Processing." *Medium*. August 20, 2023. <https://medium.com/@deepthi.sudharsan/decoding-the-relationship-language-models-and-natural-language-processing-bbc0cd6754e2>.
- Thiel, David. "Identifying and Eliminating CSAM in Generative ML Training Data and Models." *Stanford Internet Observatory*, 2023. https://stacks.stanford.edu/file/druid:kh752sm9123/ml_training_data_csam_report-2023-12-23.pdf.
- Thorn. "How Hashing and Matching Can Help Prevent Revictimization." August 24, 2023. <https://www.thorn.org/blog/hashing-detect-child-sex-abuse-imagery/#:~:text=It%20converts%20a%20piece%20of,without%20ever%20seeing%20users%20content>.
- Thorn. "We're on a Mission to Eliminate CSAM from the Internet." N.d. Accessed August 2024. <https://safer.io/about/>.
- Tidy, Joe. "Dead in 6 Hours: How Nigerian Sextortion Scammers Targeted my Son." *BBC*. June 9, 2024. <https://www.bbc.com/news/articles/c2llzppyx05o>.
- TikTok. "Edited Media and AI-Generated Content (AIGC)." *TikTok Guidelines*. May 17, 2024. <https://www.tiktok.com/community-guidelines/en/integrity-authenticity/#3>.
- Tiku, Nitasha & Tatum Hunter. "Fooled by AI? These Firms Sell Deepfake Detection that's 'REAL 100%.'" *The Washington Post*. May 12, 2024. https://www.washingtonpost.com/technology/2024/05/12/ai-deepfakes-detection-industry/?mc_cid=0e5a81bd72&mc_eid=736b3bbd0f.
- U.K. Department for Science, Innovation & Technology. "What the Online Safety Act Does." *Gov.uk*. May 8, 2024. <https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer#:~:text=The%20Online%20Safety%20Act%202023,users%20safety%20on%20their%20platforms>.
- UNICEF. "Ending Online Child Sexual Exploitation and Abuse." 2021. <https://www.unicef.org/media/113731/file/Ending%20Online%20Sexual%20Exploitation%20and%20Abuse.pdf>.
- United Nations and UN Human Rights. "Sale and Sexual Exploitation of Children." UN, 2023. <https://www.ohchr.org/en/documents/thematic-reports/a78137-sale-sexual-exploitation-and-sexual-abuse-children-report-special>.
- United Nations Interregional Crime and Justice Research Institute. "Introduction: Responsible AI Innovation." UNICRI, 2024. https://unicri.it/sites/default/files/2024-02/01_Intro_Resp_AI_Innovation_Feb24.pdf.
- United States Attorney's Office, Western District of North Carolina. "Charlotte Child Psychiatrist Is Sentenced to 40 Years in Prison for Sexual Exploitation of a Minor and Using Artificial Intelligence to Create Child Pornography Images of Minors." *Press Release*. November 8, 2023. <https://www.justice.gov/usao-wdnc/pr/charlotte-child-psychiatrist-sentenced-40-years-prison-sexual-exploitation-minor-and>.
- Warren, Scott & Joseph Grasser. "Japan's New Draft Guidelines on AI and Copyright: Is It Really OK to Train AI Using Pirated Materials?" *Privacy World*. March 12, 2024. <https://www.privacyworld.blog/2024/03/japans-new-draft-guidelines-on-ai-and-copyright-is-it-really-ok-to-train-ai-using-pirated-materials/>.
- The White House. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. 2023. <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>.
- Whitson, Rhiana. "Principals Say Parents Need to be Vigilant as Explicit AI Deepfakes Become More Easily Accessible to Students." *ABC News*. June 25, 2024. <https://www.abc.net.au/news/2024-06-25/explicit-ai-deepfakes-students-bacchus-marsh-grammar/104016178>.
- Woollacott, Emma. "Tech Firms Pledge to Eliminate AI-Generated CSAM." *Forbes*. April 24, 2024. <https://www.forbes.com/sites/emmawoollacott/2024/04/24/tech-firms-pledge-to-eliminate-ai-generated-csam/>.
- Worldometer. "GDP by Country." Accessed May 20, 2024. <https://www.worldometers.info/gdp/gdp-by-country/>.
- Zewe, Adam. "Explained: Generative AI." *MIT News*, Massachusetts Institute of Technology, November 9, 2023. <https://news.mit.edu/2023/explained-generative-ai-1109...>





IMPRINT

Authors

Bracket Foundation
8560 W. Sunset Blvd, Suite 700
West Hollywood, CA 90069, USA
info@bracketfoundation.org

UNICRI Centre for AI & Robotics
Alexanderveld 5, 2585
The Hague, Netherlands
unicri.aicentre@un.org

Clara Péron, Lisa Maddox, & Lana Apple
Value for Good GmbH
Französische Str. 47, 10117
Berlin, Germany
mail@valueforgood.com

Funding Partner

Bracket Capital
8560 W. Sunset Blvd, Suite 700
West Hollywood, CA 90069, USA
info@bracketcapital.org

Design, Layout and Illustrations

Malena Baum
secret monkey Design
malena@secret-monkey.com

Copyright © 2024 Bracket Foundation. All rights reserved.
This publication or any portion thereof may not be
reproduced in any manner whatsoever without the
express written permission of the publisher.