**Legislative Council Staff**

*Nonpartisan Services for Colorado's Legislature*

# *Memorandum*

January 2, 2026

**TO:** Interested Persons

**FROM:** Dhivahari Vivek, Science and Technology Policy Program Fellow

**SUBJECT:** Deepfakes and AI-Generated Intimate Images Involving Minors

## Summary

This memorandum reviews key federal and state laws that regulate generative artificial intelligence (AI) with particular attention to deepfakes and synthetic pornography involving minors. It describes how visual generative AI models work, including the methods used to generate AI content and realistic human imagery, and explains the legal frameworks designed to prevent misuse, especially when such content impersonates individuals or produces nonconsensual explicit material of minors.

## Deepfakes and Synthetic Visual Content

Generative AI can be used to create both deepfakes and entirely synthetic content featuring human beings. ***Deepfakes*** manipulate existing images or videos of real people to portray them doing or saying things they did not actually do or say. ***Synthetic content*** features individuals entirely generated by AI—people that do not exist in real life—doing or saying whatever the content creator programs them to do or say. This distinction can also be made regarding deepfakes versus synthetic content in the context of child exploitation. Sexually exploitative deepfakes of children include an actual child's identifying characteristics on a body other than their own performing a sexual act that did not occur, while synthetic pornography involves AI-generated children that do not exist engaging in sexual activities.

## How AI Visual Media is Generated

Generative AI can produce child sexual abuse material (CSAM) in two ways: either a model's training dataset contains CSAM directly, or it includes adult pornography with which the model can combine images of children.[1] While there are a number of methods that can be used to create visual generative AI models, this memo will describe two commonly used machine learning models in visual media generation: diffusion models and general adversarial networks (GANs).

## Diffusion Models

Diffusion models are at the forefront of visual generative AI, most notably utilized by popular text-to-image models like OpenAI's DALL-E and Stability AI's Stable Diffusion. At a high level, diffusion models generate high-quality media by first gradually modifying an authentic image and then reversing that process to reconstruct the original image.[2] Inspired by physics, the development of diffusion models that include text-to-image features involves three processes: forward diffusion, reverse diffusion, and conditional diffusion.

- **Forward diffusion.** In forward diffusion, the model, over many cycles, incrementally adds "noise" to a training image until the image starts to lose its features and become unrecognizable. To start, a diffusion model generating images of humans will be initially fed a real image of a human. During each cycle of the forward diffusion process, every pixel of this image (which is represented digitally as a set of numbers representing color and transparency) will be incrementally changed by a controlled random amount (called "noise"). For instance, if a pixel in the image was red, the next cycle might randomly modify this pixel to become very slightly less red and towards a different color, like green. The changes to each pixel during each cycle begin very small and get progressively larger, and this process is repeated thousands of times until the image has lost its original form and structure. The shapes and edges in the image become more and more blurred, and the original image is transformed to one that can often resemble TV static.
- **Reverse diffusion.** Next is reverse diffusion, where the now unrecognizable image is gradually restored to resemble the original image of the individual. Through many iterations of structured and controlled steps, the model essentially learns how to backtrace its steps to change each pixel's modified color back to its original color. Through this process, the model learns how to remove noise and detect structured patterns within the image's data to reveal more features of the image, such as being able to detect an arm or specific facial features. This process results in the model's eventual reconstruction resembling the original image of the human.

[1] David Thiel, Identifying and Eliminating CSAM in Generative ML Training Data and Models, 2023.

[2] For more information about diffusion models, see this video from IBM Technology.

- **Conditional diffusion.** Popular tools that use text-to-image features also involve conditional diffusion, where the reverse diffusion process is guided by text prompts entered by users of the AI tool. How the model backtraces its steps to return to the original image, and what patterns in the image to detect and remove, will also take into account the text given by the user.

## General Adversarial Networks (GANs)

Publicly used tools like StyleGAN, CycleGAN, and BigGAN utilize general adversarial networks, or GANs. GANs are an older but still widely used model to create image and video deepfakes and synthetic material, though are less efficient to train compared to diffusion models. GANs are trained by two models: a generator model and a discriminator model.[3] The discriminator model is first trained by being fed a large collection of real images and told to analyze the attributes that make up those real images. In a GAN generating images of people, the discriminator model would theoretically need a large volume of photos of real children to understand the features it needs to correctly recognize real children (such as facial features). The discriminator is also fed images not of humans, so that this model knows when to determine that an image does not contain a human.

Once the discriminator becomes good at recognizing real human faces, the generator model will randomly generate a fake image. This fake image is sent to the discriminator which then decides if the generator's image is a real image of a human. Both the generator and discriminator receive the feedback (real or fake). If the discriminator successfully spots the fake, the generator changes its model to produce a better fake image. If the discriminator fails, the discriminator changes its model to be better at identifying fake images. This back-and-forth between the discriminator and generator continues until the generator produces such realistic fake images that the discriminator can no longer identify them as fake.

For a high-quality GAN to produce realistic synthetic images of humans, the model would theoretically require between 50,000 to 100,000 images. However, successfully training GANs requires the training data to have wide facial variability, such as having faces at different angles and in different lighting.[4] Even if the model's developer did not have access to thousands of images, the GAN could still be effectively trained with access to fewer photos using methods like

[3] For more information about GANs, see this video from IBM Technology.

[4] Simranjeet Singh, et al., Using GANs to Synthesize Minimum Training Data for Deepfake Generation, 2020.

data augmentation—including flipped, cropped, blurred, color-distorted, or rotated versions of real images to better train the model at recognizing and generating images.[5]

## Eliminating CSAM from Image Generating Models

More recent visual generative models, including tools like Stable Diffusion 1.5 and Google's Imagen, were trained on publicly available datasets of millions to billions of images scraped from the internet (e.g., LAION-5B and LAION-400M). In 2023, the Stanford Internet Observatory found that the LAION-5B dataset included a significant amount of pornographic imagery, with at least 3,226 entries linking to images of suspected CSAM.[6]

There are challenges with removing CSAM from existing image datasets. Effectively filtering and removing CSAM is difficult when the data is systematically scraped from the internet, and may require developers to illegally access CSAM.[7] Existing image filtering techniques can screen out some CSAM, but have limitations. Challenges include the subjective nature of CSAM image labels (which may require domain expertise from child safety experts); concerns with filtering too little data, which may still leave CSAM in the training dataset; as well as concerns with filtering too much data, which could impact the quality of the model's image outputs.[8]

When compiling image training datasets prior to training the image generating model, developers could check the images against known lists of CSAM. Developers could explicitly filter content from websites known to host CSAM, or work with organizations like the National Center for Missing & Exploited Children (NCMEC) which use tools like Microsoft's PhotoDNA to create and store lists of "fingerprints" (called hashes) of previously identified CSAM images, through a process known as perceptual hashing.[9] Developers could submit their datasets to these organizations that could create image fingerprints for that dataset and compare them to those of CSAM images to determine image similarity and potentially identify any CSAM present in the model's training dataset. However, methods like perceptual hashing require access to the actual image data. For datasets like LAION-5B, which did not contain the actual images but rather the links to images, image entries might no longer have active URLs, or the websites

[5] Tang, et al., Explaining the Effect of Data Augmentation on Image Classification Tasks, 2022.

[6] David Thiel, Identifying and Eliminating CSAM in Generative ML Training Data and Models, 2023.

[7] National Institute of Standards and Technology (N.I.S.T.), Reducing Risks Posed by Synthetic Content, 2024.

[8] National Institute of Standards and Technology (N.I.S.T.), Reducing Risks Posed by Synthetic Content, 2024.

[9] Microsoft, PhotoDNA.

themselves may have taken down the images.[10] Additionally, perceptual hashing may be unable to detect "new" or previously unreported CSAM. Developers could also train future models on vetted data, such as licensed stock images and data in the public domain; however, this may be costly and may not be sufficient for training larger diffusion models.[11]

It is also difficult and questionably effective to remove these images once datasets containing CSAM have already trained image generative models, including tools that are still in use today. While the images could be removed from the original websites, instances of CSAM can still remain in downloaded copies of the original dataset that are in the possession of researchers and AI developers. Additionally, it is incredibly difficult to retrain an image generative model with a "cleaned" dataset. There are proposed methods to train models to partially discard specified concepts without re-training the model from scratch.[12] These methods may be effective, but could have impacts on the model's ability to generate non-illegal images and may be too infeasible for developers to perform.[13]

## The Role of AI in Child Victimization

The NCMEC's CyberTipline observed a 1,325 percent increase in reports involving generative AI between 2023 and 2024, totaling nearly 67,000 reports in 2024. In just the first six months of 2025, NCMEC received 440,000 AI-related reports—a more than 550 percent increase from the entire previous year's report counts.[14]

In addition to the creation of deepfake CSAM, AI can pose other risks to children:

- **Online enticement.** An individual intent on committing a sexual offense can use AI tools to create fake accounts on social media to communicate with a child.
- **Sextortion.** Individuals can use AI to create deepfake CSAM with the intent of blackmailing a child for additional sexual content.
- **Bullying and peer victimization.** As AI technology is increasingly easily accessible, children can access AI tools to create images, such as intimate deepfakes of fellow classmates, and circulate them.

[10] David Thiel, Identifying and Eliminating CSAM in Generative ML Training Data and Models, 2023.

[11] National Institute of Standards and Technology (N.I.S.T.), Reducing Risks Posed by Synthetic Content, 2024.

[12] Nupur et al., Ablating Concepts in Text-to-Image Diffusion Models, 2023.

[13] David Thiel, Identifying and Eliminating CSAM in Generative ML Training Data and Models, 2023.

[14] National Center for Missing & Exploited Children, CyberTipline Report 2024, and preliminary 2025 data.

## Federal Law

Federal law addresses nonconsensual deepfakes and AI content of minors primarily through the TAKE IT DOWN Act, which criminalizes the nonconsensual publication of intimate images, including those created with AI, and creates requirements for social media platforms and online sites.

### TAKE IT DOWN Act

The Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act, or the TAKE IT DOWN Act[15], prohibits nonconsensual publication of both real and deepfake intimate images to social media platforms and online sites. The law seeks to address loopholes in existing laws like the Digital Millennium Copyright Act (DMCA), where victims do not own the copyright to AI-generated images and thus are unable to take down those images if published online without their consent.

Under the TAKE IT DOWN Act, by May 19, 2026, online platforms are required to clearly communicate how individuals can submit requests to take down nonconsensually shared intimate images. These companies have 48 hours to take down the original violating image and make reasonable efforts to identify and remove all instances of that image. Covered platforms generally include social media services, and do not include email services, internet service providers, and online websites where content is not user generated. Companies that fail to comply with these requirements are subject to the penalties outlined in the Federal Trade Commission Act (FTCA) for committing an unfair to deceptive act or practice.

Anyone who shares or threatens to share an authentic intimate image of a minor, or shares a deepfake of a minor, may be fined and face up to three years of imprisonment. Anyone who threatens to share a deepfake of a minor may be fined and face up to 30 months of imprisonment.

## State Laws

All 50 states, the District of Columbia, and two U.S. territories have some form of protection against nonconsensual disclosure of intimate imagery. These laws generally include distinct penalties for offenses involving images of minors. Since 2019, states have also adopted legislation targeting the use of deepfakes.

[15] Public Law No. 119-12.

The 2024 and 2025 legislative sessions saw a notable increase in the passage of state deepfake laws. According to the National Conference of State Legislatures (NCSL) at least 97 state laws passed within these two years. As of 2025, over 40 states have enacted legislation relating to sexually explicit deepfakes or computer-generated imagery. Many states have expanded existing "revenge porn" laws and child sex crimes statutes, either by explicitly creating new definitions for computer-generated images and deepfakes[16] or by expanding existing statutory definitions to include images generated by AI or by computer-generated methods.[17] In contrast, states like Louisiana have adopted standalone laws related to deepfakes and non-consensual intimate images.[18]

According to NCSL, at least 29 states have enacted laws pertaining to deepfake or AI-generated CSAM. Of these, at least 12 states specifically address artificially generated synthetic child pornography, or AI CSAM that does not depict real children. For example, states like Alabama,[19] Arizona,[20] and Nebraska[21] include in their statutory definitions of CSAM computer-generated images or video featuring someone that a "reasonable" or "ordinary" person would conclude is of an actual child. Minnesota criminalizes CSAM that includes depictions of individuals "indistinguishable" from actual minors.[22] California law covers both visual depictions of children as well as AI-generated depictions of what may appear to be a child.[23]

[16] North Dakota House Bill 1386 (2025)

[17] California Senate Bill 1381 (2024)

[18] 14 La. Rev. Stat. Ann. § 73.13-14.

[19] Ala. Code § 13A-12-190, *et seq*.

[20] Ariz. Rev. Stat. § 13-3551.

[21] Neb. Rev. Stat. 28-1802.

[22] Minn. Stat. § 617.246 thru 247.

[23] Cal. Penal Code § 311.

## Colorado Law

Colorado law provides both civil and criminal penalties for the nonconsensual creation and distribution of intimate deepfakes.

- **Criminal provisions.** Senate Bill 25-288 changed existing crimes related to nonconsensual disclosure of intimate images by including computer-generated images, and created a class 6 felony if the disclosure posed an imminent and serious threat to the depicted person’s or immediate family’s safety. The bill established exceptions to liability, outlining that the crime does not apply if disclosures were made in good faith to law enforcement while reporting a crime, or to the courts and fact finders involved in a criminal proceeding. For cases involving sexual exploitation of a child, the law now eliminates the requirement for prosecutors to establish the identity of an alleged victim. The bill also expanded the definition of "sexually exploitative material" to include realistic computer-generated depictions, especially concerning CSAM. Colorado does not criminalize synthetic child images. AI-generated child sex abuse material must depict an “identifiable child,” defined as a child who is identifiable either by their face or by a distinguishing feature (like a birthmark).
- **Civil action.** Victims of nonconsensual intimate deepfakes may bring a civil lawsuit against the perpetrator, and may recover actual damages (including emotional distress) or $150,000, whichever is greater, plus attorney fees.

Colorado also has broader AI regulation. Effective June 30, 2026, Senate Bill 24-205 focuses on consumer protections and preventing "algorithmic discrimination" in high-risk AI systems (e.g., in employment, education, or housing decisions).[24] This law does not specifically target deepfakes but regulates the use of AI more generally.

## Issues for Further Consideration

Federal and state laws regarding AI intimate images, especially laws criminalizing synthetic intimate images, raise several concerns, including:

- **Lack of a “real” human victim.** The absence of a “real” victim in synthetic pornography, especially involving children, may raise both legal concerns regarding successful prosecution as well as investigatory issues. There may be debate around who specifically is harmed when CSAM features AI-generated children that don’t exist. Additionally, investigating depictions

[24] Section 6-1-1701, *et seq.*, C.R.S.

of synthetic children may create challenges for law enforcement, who often depend on the real identity or identifiable markers of a real individual to determine their age.

- **Determining intent and liability.** In addition to the challenges discussed regarding the development of image generating tools, there are ongoing discussions regarding the intent and liability of those that deploy, or use, these models. Because commonly used tools relied on datasets like LAION-5B with significant amounts of CSAM present, the repercussions of the training process for models like Stable Diffusion 1.5 will likely remain.[25] It is plausible that these models may continue to create synthetic CSAM, either because CSAM present in the training data directly influences the model's understanding of a depiction of a child, or by combining prompts (such as "child" and an explicit act). For these models, it is theoretically possible that objectionable images could be created without specific intent to do so.
- **First amendment speech considerations.** It is unclear whether broader laws prohibiting synthetic explicit material may trigger First Amendment free speech concerns.[26] The 2001 U.S. Supreme Court decision *Ashcroft v. Free Speech Coalition* struck down two provisions of the federal Child Pornography Act of 1996 criminalizing visual depictions that appeared to be, or conveyed the impression of, minors engaging in in sexually explicit content.[27] The court concluded that the federal law was overly broad, prohibiting speech that was not obscene or child pornography, and cited as examples performances of *Romeo and Juliet* and movies like *American Beauty*, which include visual depictions of teenagers engaging in sexual activity.

The federal government and states will continue to grapple with these issues and others as AI technology rapidly evolves, implementation challenges arise, and new case law emerges.

[25] David Thiel, Identifying and Eliminating CSAM in Generative ML Training Data and Models, 2023.

[26] Harshita K Ganesh, Protecting Children Through Deepfake Child Pornography: A Moral, Legal, and Philosophical Discussion on the Intersection of the Evolution in Law and Technology, 2022.

[27] *Ashcroft v. Free Speech Coalition*, 535 U.S. 234 (2002)