Child Protection and Practice 2 (2024) 100043

Contents lists available at ScienceDirect

Child Protection and Practice

journal homepage: www.sciencedirect.com/journal/child-protection-and-practice

Artificial intelligence and CSEM - A research agenda

Chad M.S. Steel

George Mason University, Fairfax, VA, USA

1. Introduction

In the near-future portrayed in the movie "The Artifice Girl", advances in artificial intelligence allow law enforcement to create a fully realized avatar depicting a 12-year-old girl whose visual and audio representations are indistinguishable from a real person. The avatar is capable of thousands of simultaneous conversations in which potential child sex offenders are caught attempting to solicit sexual activities from her, believing her to be a real person (Ritch, 2022). While the movie is science fiction, many of the enabling technologies are present today - Large Language Model[1] (LLM)-based chatbots, artificial intelligence (AI) deepfakes, and domain-specific avatars are all possible (though not yet with the real time responsiveness and fidelity depicted in the film). While the film portrays AI as being a boon to law enforcement, real world online child sexual exploitation is closer to an arms race, with technologies that can enhance detection, response, and treatment competing with those that can create content, automate grooming, and create previously unexplored legal situations.

Artificial intelligence is a broad category of technologies that attempt to model human thought and behaviors using computer algorithms (Sweeney, 2003). In the context of this research, two primary applications of AI are considered - classifiers and content generators. Classifiers are computer programs that are trained on a large dataset of content (either labeled or unlabeled), then used to categorize previously unknown content. Content generators, or generative AI, are likewise trained on a large dataset (generally using millions or billions of pieces of content) and allow for the creation of text, images, or videos from text prompts. While other approaches are referenced, the majority of current AI research related to CSEM makes use of artificial neural networks (ANNs), a form of AI that can "learn" to identify non-obvious connections between items (classification), and to generate novel combinations of items (e.g., elements of a picture) based on the training data used.

Tools like ChatGPT and DALL-E use sophisticated variants of ANNs and are trained on billions of documents, pictures, and videos extracted from the Internet (DALL·E 3, 2024). ChatGPT, one of the most prominent tools for textual AI generation, was trained using books, websites, and articles to be able to generate similar content using a prediction-based approach. For example, if the first words of a sentence are "The quick brown", ChatGPT will predict "fox" as the most likely next word based on examples from its training data. Additionally, ChatGPT's developers trained an instruction parser to understand prompts (e.g., "Write me a poem about lilacs in the style of Keats") as input to its prediction engine (Johri, 2023). Similarly, DALL-E was trained on millions of images with associated captions (Betker et al., 2023), and uses a similar instruction parser to generate images from text prompts (e.g., "Produce a picture of a swan in the style of Monet"). Both technologies are hosted by OpenAI and have guardrails incorporated into them - ChatGPT, for example, prevents prompts for generating malicious code or phishing scams (Alotaibi et al., 2024), while DALL-E, excluded sexually explicit images from its training data (OpenAI, 2022). Despite these guardrails, researchers have found ways to bypass the controls (Alotaibi et al., 2024), and criminals have created their own, similarly designed generative AI tools that are both intentionally trained on problematic data and remove the front-end prompt controls (Falade, 2023).

The struggle between technology both enabling and helping combat online criminality has been present for Child Sexual Exploitation Material (CSEM) offending since the inception of networking, with offenders adopting new technologies (Steel et al., 2020) at the same time as others attempt to use them for deterrence and detection (e.g., Quayle, 2020). Now that readily available AI, particularly that based on deep learning driven ANNs (ANNs with multiple layers) has become computationally feasible and readily available through tools like Gemini and Stable Diffusion, there is an immediate need for additional research into the intersection of these technologies and online criminality. Deepfake technologies have already been used extensively by criminals for everything from state-sponsored disinformation campaigns (Whyte, 2020) to creating sound-alike audio of relatives in distress for use in phone fraud scams (Audrey & Smaili, 2022). CSEM offenders have begun to use these technologies as well. Tools like Undress AI allow for the de-aging and undressing of individuals depicted in innocent images and their use is rising exponentially (Murphy, 2023). Meanwhile, research into their usage by criminals and their operational usage by law enforcement for detection and deterrence efforts has lagged. As such, there is a critical need for rapid and robust research into the use of AI

[1] Large Language Models are a type of AI that are trained on a very large corpus of text to allow them to predict and generate new, plausible language based on user prompts.

https://doi.org/10.1016/j.chipro.2024.100043
Received 9 March 2024; Received in revised form 27 May 2024; Accepted 29 May 2024

related to CSEM from both an offensive and defensive perspective.

Many AI technologies currently operate in an uncertain policy and legal environment within the United States as well. The original intent behind child pornography legislation was the protection of real children from physical abuse (Ashcroft v. American Civil Liberties Union, 2004), though new legal theories need to be applied with advancements in AI. Issues include laws around unauthorized (or even consenting) de-aging, misuse of an individual's innocent images (or audio) to generate offending content, and the creation/possession of fully synthetic content. The public's support for this content being illegal is high, however, with 81% believing virtual CSEM should be illegal (Steel et al., 2022b), which is encouraging for potential legislative fixes.

This paper highlights recent AI advances and limitations and their intersection with CSEM offending in three key areas - content detection; online grooming and social CSEM offending; and generative AI content creation. Other recent research has been published on AI related to CSEM by Singh and Nambiar (Singh & Nambiar, 2024), who performed a systematic review of AI algorithms focused exclusively on the prevention of CSEM primarily from a technical information security perspective. That work identified current trends in AI prevention, as well as recommendations for algorithm evaluation and usage. Additionally, AI has been applied to other areas of CSEM, including the use of natural language processing to extract case details for risk assessment purposes (Cohen, 2023). This paper builds on those and other works highlighted below, looking at both offensive and defensive uses of AI. Key current research from each area are highlighted. Based on the current research and trend information, a series of research gaps and suggested areas for improvement are then presented to facilitate the creation of policy and research agendas to combat this threat.

2. Content detection

One of the most active areas of child sexual exploitation research is the automated detection of CSEM (for a more comprehensive survey of CSEM detection in general, see H.-E. Lee et al., 2020). Established detection techniques have historically relied on hash matching. In hash matching, databases of unique digital signatures (hashes) from previously seen offending content are utilized. Traditional file-based hashes (e.g., SHA-256) have been used by organizations such as the National Center for Missing and Exploited Children (NCMEC) and the Internet Watch Foundation (IWF) to create hashsets of this material, which are then utilized by content providers and forensics examiners to identify offending content. Traditional hashes only allow for the identification of exact content matches - any modifications, including resizing, cropping, or saving in a different format will result in completely different signatures. To address this, these organizations also use content-driven hash techniques, including PhotoDNA, to improve detection. Unlike traditional hashing, PhotoDNA generates a more robust signature that allows for the identification of images that have been cropped, resized, or saved into a different image file format (Microsoft, 2009). While hash techniques are effective at identifying previously known CSEM, they can't be used to identify new content (nor highly modified content) and both forensic practitioners and network providers have identified a need for more advanced tools (Sanchez et al., 2019).

Early CSEM content detection efforts to identify previously unknown content focused primarily on feature-based identification but suffered from limitations related to both precision and recall (Gangwar et al., 2017). Newer AI-based research has utilized various techniques that have greatly advanced the accuracy over earlier feature-detection approaches, many using deep learning through ANN variants, to identify CSEM. Image-based CSEM detection faces two primary tasks - identifying pornographic (or explicit) images, and determining the age of the individuals portrayed. Pornographic image detection using AI is a relatively mature research area (see Cifuentes et al., 2022)), but AI CSEM detection has the added challenge of age determination. Historically, these were combined into cascading classifiers (the use of multiple classifiers in series), with advances in the individual techniques incorporated as discrete steps using a pipeline approach (Sae-Bae et al., 2014).

Recent advances in discrete image-based CSEM detection have shown feasibility for research purposes. Gangwar et al. (2021) utilized a specialized ANN that was trained on a corpus of adult SEM in conjunction with manually labeled non-offending images of children, treating the problem as a two-stage identification, then tested their tool on a police-provided corpus of 5,000 actual CSEM images, reporting a 93% accuracy. For age identification, Anda et al. utilized an ANN-based deep learning model called DeepUAge to improve upon these techniques. Additionally, they developed a labeled facial age corpus titled VisAGe for use in developing and comparing other models (Anda et al., 2020). Utilizing ANN-based facial age estimation and building upon Yahoo!'s NSFW explicit image detection approach (Woodie, 2016), Rondeau (2019) developed a similar pipeline-based CSEM classifier, which yielded accuracy in the 80–90% range when tested on a real world dataset. Looking at non-consensual sharing of nude images, SafeSext, a newer proof-of-concept messaging application, utilizes a network that was trained to identify potentially problematic images based on AI-based fingerprinting (Franco et al., 2024).

When treated independently (as many early classifiers did), the error rates from both stages become multiplicative, resulting in both high false positive and false negative rates that are unacceptable in large scale applications (e.g., processing millions of images a day as providers such as Meta or Google are required to do). These approaches were potentially viable for forensic triage on smaller datasets, but have been found to be problematic due to the error rates on larger dataset applications due to the base rate problem (Dalins et al., 2018). Additionally, these discrete approaches fail to incorporate context clues that human examiners might use to better discriminate in CSEM identification, focusing exclusively on victim developmental identification and the presence of unclothed body parts (which may not be sufficient nor dispositive for CSEM) (Kloess et al., 2019).

Newer approaches have used deep ANN architectures to combine both individual classifiers into a single solution, allowing AI to make the appropriate age-related inferences. Combined AI-based detectors have multiple additional advantages over separated detection in that they can incorporate additional features not available in face recognition (e.g., genitalia development) and may be less dependent on camera angle. AI-based detectors may also implicitly incorporate contextual clues (e.g., clothing or environment) that may be more indicative of CSEM content - a child's room, for example, may have different items and environmental design than an adult's room (Laranjeira da Silva et al., 2022). Vitorino et al. (2018) attempted to utilize a single tier classifier with an ANN for CSEM detection with accuracy in the mid-80 percentile range, but ultimately had slightly better success with a cascading classifier due to the size limits of an available training dataset for CSEM. Recent advances in computing and iterations of ANN-based architectures are expected to improve these numbers substantially based on similar improvements identified in traditional SEM detection (Cifuentes et al., 2022).

In addition to the approaches adopted from adult SEM and environment-based approaches, others have used various language models to identify CSEM content based on file naming. CSEM offenders use domain-specific terms that are reflected in metadata such as filename conventions that are distinct from other content and specific to a particular technology - e.g., web-based search terms differ from those on peer-to-peer networks (Panchenko, Beaufort, & Fairon, 2012; Steel, 2014a). The use of unusual terms with unusual parsing characteristics (e.g., "R@yGold"), as well as the lack of a large enough corpus, limits the applicability of LLM-based approaches trained on general language datasets, but success has been shown with feature models that use multi-word phrases for context (Peersman et al., 2016). This has been extended to a forensics context through the use of both filenames and file paths using both general machine learning (ML) and ANN-based classifiers (Al-Nabki et al., 2023), including using adversarial manipulation of filenames/paths to improve generalizability (Pereira et al., 2020).

Training these applications can be difficult as well. In the United States and many other countries, possession of child pornographic material, even for research purposes, is illegal, though there are limited datasets available in other countries, such as Brazil's Region-based annotated Child Pornography Dataset (RCPD), which contains approximately 2000 CSEM images (Macedo et al., 2018). Additionally, a file path dataset available from ProjectVIC has been used with short text classifiers (Pereira et al., 2020). Because of the limited general datasets available, many of the approaches used are tested only on synthetic data, and efforts to create better synthetic data with the appropriate features have been proposed (Yiallourou et al., 2017), but how accurately they reflect real CSEM is unknown.

3. Online grooming and social CSEM

A subset of CSEM transactions is conducted by individuals who have a more social predilection, which can include everything from one-to-one communications (e.g., instant messaging) to many-to-many communications (e.g., posting to online forums), and involve communications with either victims or other offenders. Because online grooming of children can also involve both the transmission of existing content as well as the use of coercion/extortion to produce new content, these topics are considered together for the purposes of this paper (Steel, 2021).

Machine learning-based analysis in online forums (in this case, the dark web) has been shown to be effective in identifying the topicality of forums using clustering techniques (i.e., grouping similar content together) based on phrases within the postings to categorize their content (Nazah et al., 2021). Ngo et al., (2023) successfully used a labeled dataset of CSEM and non-CSEM posts to dark web forums and applied a combination of traditional classifiers and ANNs that took a holistic approach in looking at the totality of content in each post in their classification work.

One of the most recent approaches to identifying child sexual exploitation-focused conversations has utilized current generation LLMs and ANNs. The highest performing of these, presented by Borj et al. (2023), used an optimized ANN (Liu et al., 2019), with a traditional classifier to obtain a 99% accuracy on the PAN 2012 dataset (Inches & Crestani, 2012). From an LLM perspective, an upcoming model using Llama (a specific LLM implementation) showed a >98% accuracy in identifying grooming messaging (Nguyen et al., 2023) using a more realistic combination of the PAN 2012 dataset and a larger more general dataset. One of the newest AI models in detecting child grooming, DRAGON-Spotter, leverages discourse analysis utilizing deep learning to detect less obvious chats. Unique in this space (and important for law enforcement purposes), DRAGON-Spotter provides explainability and specificity for potentially offending chat segments (Lorenzo-Dus et al., 2023).

A different approach to detection, similar to the movie plot noted in the Introduction, has leveraged LLMs to detect problematic behavior through chatbots. The Sweetie 2.0 chatbot, for instance, identifies offenders who engage with a virtual avatar and exhibit grooming behavior, and are then referred to police (Henseler & de Wolf, 2019). CSEM-specific chatbots either impersonating children or impersonating other offenders to obtain voluntarily shared materials are technologically feasible, but concerns about entrapment and unintended consequences limit their widespread usage in many jurisdictions. Additionally, sophisticated LLM-based chatbots that can impersonate children such as Microsoft's Tay have been possible for over a decade but found to be generally problematic (Vorsino, 2021), and they represent a dual-edged sword for CSEM offenses. They may be used to generate “realistic” responses to victims by offenders, or to engage more accurately when impersonating children when used by investigators. The blending of this technology with real-time voice and video generation will likely lead to their use by offenders for offline offending interactions as technology evolves. Finally, non-AI chatbots such as reThink have been used to engage with offenders seeking CSEM images as part of a warning/education strategy tested on Pornhub for individuals searching for offending terms (Internet Watch, 2024), and AI-enhancement is a logical next step for this use model. The initial implementation of reThink, including warning messaging and the chatbot, showed desistance as the primary outcome for individuals viewing even a single deterrence message, and a substantial number of help-seeking web referrals following interaction with the chatbot (Scanlan et al., 2024).

Similar to the classifiers attempting to identify CSEM from image and video content, there are no single, comprehensive datasets for use in text classification. Most frequently, transcripts from organizations such as Perverted Justice, who interact with alleged child sex offenders then post the chats online, have been used for classification purposes (Pendar, 2007). One of the more common datasets used in evaluating text classifiers is the PAN 2012 dataset, a collection of conversations between offenders and individuals pretending to be offenders (also derived from the Perverted Justice data) (Inches & Crestani, 2012). Faraz (2023) created an extended PAN 2012 dataset containing an additional 71 Perverted Justice chats for testing purposes (allowing the full PAN 2012 dataset to be used for training). Finally, PANC combined portions of the PAN 2012 and ChatCoder2 datasets (which utilized the PAN 2012 dataset) (Kontostathis et al., 2012) to better balance positive and negative dataset content (Vogt et al., 2021). All of the above datasets use pseudo-conversations (the individuals are pretending to be victims) for their true positive training and given the reflexive nature of conversations the pseudo-victims statements would likely impact the offender statements, requiring additional research to determine their applicability in real environments. Prior research has shown significant differences between the interactions of offenders and real victims, police posing as children, and vigilante decoys posing as children. These include differing chat lengths (with law enforcement chats being the shortest), as well as differences in responsiveness - law enforcement and vigilantes were more likely to ask for clarifications on implied actions (Ringenberg et al., 2021). These differences, as well as differences in chat stages related to the assessment of risk, planning for meetings, and sexual content have been highlighted as confounding factors in developing AI detectors (Ringenberg et al., 2024). Evaluations are impacted by the number (and type) of non-offending communications included and may not be directly applicable to all types of communications. The generalizability of these classifiers between platforms and communication types is not currently well characterized, and comprehensive datasets of conversations between offenders are not readily available (as opposed to victim-offender conversations as noted above).

While the current approaches are focused on forum or chat-based messaging, there has been minimal effort to apply them to one-on-one messaging using tools such as WhatsApp, Signal or Facebook Messenger. With the increase in mandatory encryption within these tools, vendors have abrogated responsibility for monitoring, though client-based detection would still be possible. Combining chat-based and rich content (image and video) based AI approaches, detection within clients without server-based monitoring of chats would be theoretically possible. While proof-of-concept, specialized messaging clients like SafeSext exist, to-date there have been no substantial client-based implementations for detecting CSEM in the major messaging platforms.

4. Generative image AI and virtual Child Sexual Exploitation Material

The creation of virtual CSEM is not new. As an example, John Stelmack, a school principal, physically cut-and-pasted the heads of children under his supervision onto nude adult bodies. The court found that the images did not meet the standard of a child being exploited (Stelmack v. State, 2010), however federal law in the United States makes an exception when the images in question become higher fidelity and are “virtually indistinguishable” from real children (Ashcroft v. Free Speech Coalition, 2002). Additionally, individuals in the past have used software tools such as Photoshop to “morph” images more seamlessly (e.g., by

pasting the face of a child onto adult pornography), but the ability to do so effectively required skill and effort (Seto & Eke, 2015). In the age of artificial intelligence, creating high quality images and videos of this nature rapidly is now possible even for offenders with lower technical ability.

The use of AI to create deepfake nude images is no longer a theoretical and becoming widespread - a former psychiatrist David Tatum was convicted of using a web-based tool (www.deepsukebe.io/en) to create nude images of children he knew (as well as secretly recording children covertly) for sexual gratification purposes (United States of America, V. David Tatum, Defendant, 2023). In more complex situations, students in Spain, South Korea, and the United States have all recently used AI to generate nude deepfakes of their classmates (McNicholas, 2023; O'Brien & Hadero, 2023). Extending this, criminals have begun using AI-generated images for sextortion - no longer needing to groom children into sending self-generated content to bootstrap their exploitation (Criminals Using A.I. to Alter Images for Sextortion Schemes, State Police Warn, 2023). The long-term impact for victims and the facilitation effect of this on contact offending is currently unknown. One of the most popular offline tools for text prompt-based image generation, Stable Diffusion, was recently found to have used previously known CSEM material in its training data, facilitating the ability of offenders to create customized, AI-generated exploitation material (Levine, 2023). Modified versions of the this and similar tools can be used to create highly customized CSEM content on-demand based on simple prompt engineering (e.g., "produce a realistic image of a ten-year-old, red headed girl in a bathtub"), which reduces the need to download content (and the likelihood of detection), but the overall impact on global trading is still unexplored. With the widespread availability of high-quality AI image manipulation tools, this problem is expected to grow and expand from images into videos, but limited tools are available at present to detect these images.

There are two primary points of detection in the generation of CSEM related AI deepfakes. First, AI tools can incorporate the detection of CSEM into the applications themselves, at both the time of upload and the time of generation. To be successful, both are required - both hash-based techniques and AI-based CSEM detection, as noted above, can be incorporated, with some potential enhancements. For uploaded images, the minor detection may need to be separable from the pornography detection - in particular, individuals such as those noted above can upload adult pornographic images and innocuous images of children to combine them. This requires two separate classifiers to detect (age and pornography). The second approach, detecting CSEM post-image generation (but before providing it to the user) can use the same techniques for general CSEM detection (Gangwar et al., 2021), with high likelihood images being reported automatically. While these can be incorporated directly into platforms like Google's Gemini or Midjourney, the ability for end users to download tools such as Stable Cascade (the next generation of Stable Diffusion) and insert their own front ends shows the need for LLM-based AI content generators to build prompt-based prevention (not just detection) directly into their core offerings.

While CSEM image identification is currently an active area of research, other generative AI areas are lacking in any substantial published research. The use of AI tools by CSEM offenders needs to be explicitly studied to identify the affordances available. For example, offenders may request an AI tool "de-age" an adult image, or they may request the generation of an "artistic" or "medical" image based on ingested, legal imagery. This provides an additional area of legal research as well - can de-aged images of consenting adults be considered child pornography, or can aged images of real children to depict them as adults be considered child pornography? As of now, this is an unsettled issue of law in the United States and may impact both offensive and defensive investigative areas (for example, if it is deemed legal investigators could create "synthetic" child pornography by de-ageing a consenting adult, though this would carry with it substantial ethical concerns).

In addition to automated detection by tool providers, the content generation training can be done on less risky datasets. In a review of the images used by Stable Diffusion in training its model, over 1,000 known CSEM images were identified (Cole, 2023). While the negligence of Stable Diffusion's use of poorly vetted data represents the extreme of irresponsibility, other tools may be trained using "legal" child erotica and adult pornography, which may collectively facilitate the generation of CSEM imagery, or offenders may utilize their own large collections of CSEM to "boost" or fine-tune existing AI training data. As noted above, building prompt-based detection into tools is needed due to the ability to run these tools offline, but pre-scanning (content detection) in training set creation is also needed. By providing training sets that minimize the ability to generate CSEM, providers may be able to minimize the downstream risk (or at least force offenders to supplement the training sets themselves), but the impact of such training set limitations on actual offender usage has yet to be evaluated in empirical studies.

Following their creation, detecting whether CSEM images are real, fully AI-generated, or deepfakes provides a further challenge to investigators. While the United States law does not require the images to be of real children for an offense to be charged, just that they are indistinguishable from real children (removing the prosecutorial need to provide authenticity), investigators still need to make a distinction as any real or deepfake images need to be treated as a victim identification problem (e.g., identifying the victim to stop further abuse) (Steel et al., 2022a). Current AI detection tools are in their infancy and have very low precision/recall statistics for practical purposes.[2] Two potential techniques have been put forth for the forensic detection of an AI image - spatial domain analysis (evaluating the visual features of the content) and frequency domain analysis (evaluating mathematical signals within the content). Marra et al. (2019) found that spatial domain artifacts were present in the generative AI images of many early generation tools, with the ability to forensically identify which specific tool created a given image in some circumstances (Yu et al., 2018). Spatial-domain approaches generally suffer when images are altered (e.g., resized, cropped, blurred), so frequency domain analyses were developed for more practical usage as many individuals and sites change image formats or recompress images. Zhang et al. (2019) found ANN-specific spectral peaks (unique frequencies), and upsampling artifacts (changes that occur when increasing the resolution of content) that were neural-network specific were further identified in AI generated imagery (Frank et al., 2020). Despite early promise, a recent analysis across AI tools and detection techniques found detection rates ranging from chance to approximately 90%, significantly lower than what is needed for practical use (Corvi et al., 2023).[3] Fortunately, many tools add meta-data which includes the name of the tool used to create the image to generated content, specifically exchangeable image file format (Exif) tags that are part of many common image file formats. As such, low-tech forensic efforts such as scanning Exif information for AI signatures are more viable in the short-term, though easy to alter.

5. Research needs

The rapid evolution of AI and its affordances have not been met with a commensurate uptick in research. While excellent research is occurring in specific areas related to CSEM offending and AI, additional areas are unexplored or underexplored. A recent policy paper by Thorn, focused on content developers and providers, is a step in the right direction (Thorn, 2024). The Thorn work highlights critical areas of prevention for tech

[2] Though general Deepfake detectors operate on content, digital forensics approaches may be able to utilize other factors, including the installation of relevant applications or visits to relevant generative websites.

[3] A 90% detection rate may be useful for small-scale forensics work for probability-based ranking purposes, but the base rate fallacy would apply to any usage at provider levels, where hundreds of millions of images a day need to be scanned.

companies in developing, deploying, and maintaining generative AI tools. Notably, they recommend industry-wide adherence to a set of guidelines, including incorporating reliable data and testing into the development of new tools; responsibly assessing tools before deployment and ensuring they include adequate prevention controls and content monitoring; and actively monitoring for misuse and better reporting of violations to NCMEC. Critically, the Thorn working group was composed of many majority generative AI providers, including Meta, Google, Microsoft, and OpenAI (Y. Lee, 2024). Despite the policy model and provider recommendations, underlying research gaps remain. Some key areas of near future research need are recommended below.

1. *Improve the detection of AI-generated content.* The current state of identifying synthetic CSEM is still developmental - detection rates are significantly lower than necessary for operational use. While the accurate identification of real victims for fully synthetic CSEM or de-aged CSEM isn't required for prosecution in the United States, one of the main investigative needs is victim identification to stop ongoing abuse. The development of better AI-detection techniques will prevent law enforcement from wasting limited resources on the identification of fully synthetic victims and continue the focus on identifying real victims of CSEM offenses (Steel et al., 2022a).
2. *Utilize the base rate problem in detection.* The base rate problem in CSEM detection appears to present an insurmountable challenge for AI-based approaches at a provider level. Because the prevalence of CSEM is so low on most legitimate social media platforms when compared to non-offending content, even a 99% detection accuracy results in unmanageable false positive rates. However, the most egregious offenders rarely have a single image and/or video associated with their accounts. If an entire account (or activity history) is treated as a sub-corpus, the more CSEM content present as a percentage the better the overall detection rate will be. Additionally, the likelihood rating that CSEM is present will increase the more offending content is present. Studies identifying offending accounts (instead of images) using AI tools by providers (in partnership with researchers) are needed and offer potential benefits in reduced review costs and better detection of the highest risk offenders. This can additionally be applied to text-based classification. Groomers, traders, and forum posters are likely to have multiple messaging interactions across multiple threads and/or individual chats, which can be analyzed in aggregate for potentially higher rates of sensitivity and specificity. Approaches using the entire corpus of an author (and linking accounts based on author identification) have shown high recall, though also high false positive rates (Kontostathis et al., 2012), however other contextual information (e.g., forum, timing, etc.) may reduce this. Finally, Bayesian approaches utilizing prior probabilities can be employed - a forum where teens chat and share images is likely to have a much higher rate of offending content than a web forum targeting gardening techniques, which can be utilized to improve precision and recall based on prior detection rates.
3. *Create larger video, text and image-based datasets that can be used to test and compare various tools in a secure manner.* The current corpii are relatively small and are not sufficiently representative (either content-wise or demographic-wise) for the operational needs of researchers. For image datasets, there is a need for innocuous, adult SEM, and CSEM imagery that is fully labeled, and which can be drawn from to represent actual use cases. Additionally, to obtain accurate comparisons across different tools the datasets must include a balanced, ground-truth labeled set of ages across the continuum from birth to 18 years old, ideally with the comparator adult dataset containing younger-looking actors (e.g., 18–24 years of age) whose ages are also accurately labeled. This will help address concerns that much of the current research may be biased based on higher accuracy rates at the fringes of the data (e.g., discriminating infant-oriented CSEM from adult pornography containing actors in their 40s) as opposed to the more complex problems (e.g., discriminating CSEM of a 16-year-old from that of a 19-year-old). There is the additional possibility of using AI generation tools to create these datasets (storing only the feature descriptors), which has the added potential advantage of improving the detection of other AI-generated imagery. Text-based datasets, including chat logs where CSEM is transacted, file names and search terms - both innocuous and problematic - are similarly needed. The limited data created from synthetic grooming chats and keyword lists is insufficient for research tool development, testing, and comparison. Similarly, labeled video datasets as well as datasets including both real and synthetic CSEM are virtually non-existent to support forward-looking research. Finally, with advances in voice and face recognition, there is an additional need for both victim and offender biometric datasets to facilitate the development and deployment of newer, multimodal detection tools (Westlake, Brewer, & Swearingen, 2022). Due to the limitations in possessing CSEM material, even for research purposes, blinded APIs, feature-descriptor datasets, and/or sandbox environments can be made available by governments to address these needs.
4. *Develop tools to better geolocate indoor images and videos.* Current victim identification efforts rely on human review of victim imagery to identify the location of the abuse. Frequently, these locations are indoors, and victim identification specialists look for things like power outlets and fixtures; clothing items; logos; food products; furniture; and other clues as to the origin of the image. Research using object recognition and lookup (e.g., using Google Reverse Image Search), as well as the use of general ANN-based AI recognition for indoor images would greatly benefit these efforts. An appropriately tagged dataset of indoor images (e.g., based on Exif information) to further this research, as well as new software tools, may help to identify child victims and prevent further abuse more rapidly.
5. *Utilize contextual information, including text, spatial, metadata, and image/video-based features for classifications.* Current classifiers tend to focus on one aspect (e.g., filenames or hashes) of content, and fail to include the full context. Research is needed into holistic classifiers that include all available information, including the ages of individuals depicted; the content of the images (explicit or non-explicit); the environmental cues present in the image; the filenames and paths; and the context in which the images are transacted (instant messaging or forum texts). Additionally, information from other channels needs to be incorporated - for example, work on including audio cues (ranging from age detection based on voice patterns to automated transcription/translation to identify relevant phrases) for video-based content is largely missing. Cross-content identification work has been minimal as well - identifying environmental or participant features in relation to other known-offending images (e.g., subject or victim face recognition) is done by analysts, but no large-scale work using these approaches has been released, though developmental tools such as BANE show promise (Westlake, Brewer, & Swearingen, 2022). Finally, approaches using filenames and pathnames (Al-Nabki et al., 2023) can benefit from spatial ML approaches, with files that are closer in path traversal to other CSEM content having potential higher a priori likelihoods of being CSEM.
6. *Rate message/chat classifiers based on both accuracy and speed of detection.* One underexplored area is the timing of potential discovery - i.e., how many messages does it take to identify predatory behavior? Borj et al., (2021) for example, eliminated conversations with fewer than 7 messages exchanged. The number of messages needed for a given transaction (e.g., trading CSEM or directly exploiting a child to produce CSEM) is largely unknown and is a competing factor in detection rates - ideally detectors will identify problematic behavior before full exploitation occurs. As such, a schema rating classifiers based on the rapidity of detection against the accuracy would be helpful in evaluation for practical implementation (e.g., curves showing detection rates as a function of the number of messages analyzed). Finally, current classifiers tend to blend both coercive and

extortive exploitation as a single class and may benefit from treating them as separate and distinct.

7. *Explore the victimization impact of new methods of CSEM creation.* The impact of CSEM trading and sharing on victims has been generally underexplored, in part due to ethical concerns and difficulties in doing so, but the extant research shows it to be significant (Cooper, 2012). While the impact on victims of traditional CSEM is underexplored (in particular non-consensually shared, self-generated images), the creation of de-aged content as well as altered CSEM content using real victims as a base is fully unexplored. This research will allow for the better identification of the treatments needs of victims and drive potential legislative and policy changes. When considering issues related to stricter laws on the new modes of AI CSEM creation, victim compensation, and offender sentencing, the voices of victims need to be included in a structured, evidence-based way.
8. *Identify differences in demographic, psychographic, and risk factors for AI using offenders.* It is currently an unanswered question whether or not CSEM offenders that use AI are significantly different from traditional online only, mixed, or contact-only offenders. How they become engaged in CSEM, how they use the new tools, and what risks for co-offending are present will drive deterrence and treatment efforts. Demographically, the population is likely to change as AI-based tools become more mainstream and easier to use and offenders with lower technophiliia and technical ability utilize them (Steel, 2014b).
9. *Explore the impact of widespread generative AI on production and trading of CSEM.* With the ability to self-generate high quality CSEM, there is the potential for a reduction in demand for new CSEM involving direct abuse. Commercial production in particular may decline or may shift to prompt-engineering based models where commercial providers produce "high quality" video before widespread consumer availability. The impact of this on contact offending rates, detection rates, and deterrence efforts (e.g., messaging when searches are conducted for offending terms) is unknown and unexplored.

With all of the research gaps, there is a critical need for interdisciplinary work. Computer scientists, data scientists, psychologists, criminologists, legal scholars, and other social scientists need to work together, and expanded cross-training of researchers is needed. The increased inclusion of technical AI topics in the social science curriculum, as well as the inclusion of human-computer interaction and criminology into digital forensics and data science coursework must be undertaken immediately to train the next generation of CSEM researchers.

6. Conclusions

AI is no longer a "future technology" and is both widely available and becoming more accessible for less technologically sophisticated users. This work summarized the current state of research, highlighting key papers providing current and potential near-term offending uses of AI by CSEM offenders. These uses are no longer speculative, but significant unanswered research questions remain as to their adoption and their impact (positive and negative) on traditional CSEM production and trading, as well as their impact on victims. Exploring these research areas in a timely manner will assist in developing new treatments and interventions, as well as inform approaches to detection, deterrence, and enforcement. Finally, regulatory and legal considerations need to take into account technological advances, and legislation must be informed by evidence-based research. Ultimately, the impact on AI on overall victimization and the balance of both defensive and offensive capabilities remains to be seen, but by aggressively pursuing research into this topic both academics and practitioners will be able to influence the course in a positive direction.

CRediT authorship contribution statement

Chad M.S. Steel: Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Al-Nabki, M. W., Fidalgo, E., Alegre, E., & Alaiz-Rodriguez, R. (2023). Short text classification approach to identify child sexual exploitation material. *Scientific Reports, 13*(1), Article 16108.

Alotaibi, L., Seher, S., & Mohammad, N. (2024). Cyberattacks using ChatGPT: Exploring malicious content generation through prompt engineering. In *2024 ASU international conference in emerging technologies for sustainability and intelligent systems (ICETSIS)* (pp. 1304–1311).

Anda, F., Le-Khac, N.-A., & Scanlon, M. (2020). DeepUAge: Improving underage age estimation accuracy to aid CSEM investigation. *Forensic Science International: Digital Investigation, 32*, Article 300921.

Ashcroft, v (2002). *Free Speech coalition*, 535 U.S. 234 (U.S. https://scholar.google.com/scholar_case?case=4016009721484982910

Ashcroft v. American Civil Liberties Union. (2004). 542 U.S. 656 (U.S. https://scholar.google.com/scholar_case?case=5352124576782659763.

Audrey, de R.-R., & Smaili, N. (2022). The unethical use of deepfakes. *Journal of Financial Crime, 30*(4), 1066–1077.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., & Others. (2023). Improving image generation with better captions. *Computer Science, 2*(3), 8. cdn. Openai. Com/papers/dall-E-3. Pdf.

Borj, P. R., Raja, K., & Bours, P. (2021). Detecting sexual predatory chats by perturbed data and balanced ensembles. *2021 international conference of the biometrics special interest group (BIOSIG)* (pp. 1–5).

Cifuentes, J., Sandoval Orozco, A. L., & García Villalba, L. J. (2022). A survey of artificial intelligence strategies for automatic detection of sexually explicit videos. *Multimedia Tools and Applications, 81*(3), 3205–3222.

Cohen, T. H. (2023). Building a risk tool for persons placed on federal post-conviction supervision for child sexual exploitation material offenses: Documenting the federal system's past, current, and future efforts. *Federal Probation, 87*, 19.

Cole, S. (2023). Largest dataset powering AI images removed after discovery of child sexual abuse material, 404 Media https://www.404media.co/laion-datasets-removed-stanford-csam-child-abuse/.

Cooper, S. W. (2012). The impact on children who have been victims of child pornography. *Written Testimony before the US Sentencing Commission*. https://www.ussc.gov/sites/default/files/pdf/amendment-process/public-hearings-and-meetings/20120215/Testimony_15_Cooper.pdf.

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., & Verdoliva, L. (2023). On the detection of synthetic images generated by diffusion models. In *Icassp 2023 - 2023 IEEE international conference on acoustics, Speech and signal processing (ICASSP)* (pp. 1–5).

Criminals using, A. (2023). *I. to alter images for sextortion schemes, state police warn.* Pittsburgh: CBS. https://www.cbsnews.com/pittsburgh/news/artificial-intelligence-alter-images-sextortion-schemes-warning/.

Dalins, J., Tyshetskiy, Y., Wilson, C., Carman, M. J., & Boudry, D. (2018). Laying foundations for effective machine learning in law enforcement. Majura – a labelling schema for child exploitation materials. *Digital Investigation, 26*, 40–54.

Dall E 3. https://openai.com/dall-e-3, (2024).

Falade, P. V. (2023). Decoding the threat landscape : ChatGPT, FraudGPT, and WormGPT in social engineering attacks. *arXiv [cs.CR]. arXiv*. http://arxiv.org/abs/2310.05595.

Faraz, A. (2023). Curated PJ Dataset. *IEEE Dataport*. http://10.21227/4kyv-n442.

Franco, M., Gaggi, O., & Palazzi, C. E. (2024). Can messaging applications prevent sexting abuse? A technology analysis. *IEEE Transactions on Mobile Computing, 23*(2), 1613–1626.

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., & Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. In H. D. Iii, & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 3247–3258). PMLR.

Gangwar, A., Fidalgo, E., Alegre, E., & González-Castro, V. (2017). *Pornography and child sexual abuse detection in image and video: A comparative evaluation* (pp. 37–42).

Gangwar, A., González-Castro, V., Alegre, E., & Fidalgo, E. (2021). AttM-CNN: Attention and metric learning based CNN for pornography, age and Child Sexual Abuse (CSA) Detection in images. *Neurocomputing, 445*, 81–104.

Henseler, H., & de Wolf, R. (2019). Sweetie 2.0 technology: Technical challenges of making the sweetie 2.0 chatbot. In S. van der Hof, I. Georgieva, B. Schermer, & B.-J. Koops (Eds.), *Sweetie 2.0: Using artificial intelligence to fight webcam child sex tourism* (pp. 113–134). T.M.C. Asser Press.

Inches, G., & Crestani, F. (2012). Overview of the international sexual predator identification competition at PAN-2012 CLEF (Online Working Notes/labs/workshop). *CLEF* (Vol. 30). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=3cf1dc2ee5a59a77ef7d633c641211a893757811.

Johri, S. (2023). The making of ChatGPT: From data to dialogue. https://sitn.hms.harvard.edu/flash/2023/the-making-of-chatgpt-from-data-to-dialogue/.

Kloess, J. A., Woodhams, J., Whittle, H., Grant, T., & Hamilton-Giachritsis, C. E. (2019). The challenges of identifying and classifying child sexual abuse material. *Sexual Abuse: A Journal of Research and Treatment, 31*(2), 173–196.

Kontostathis, A., West, W., Garron, A., Reynolds, K., & Edwards, L. (2012). *Identifying predators using ChatCoder 2.0. ceur-ws.org*. https://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-KontostathisEt2012.pdf.

Laranjeira da Silva, C., Macedo, J., Avila, S., & dos Santos, J. (2022). Seeing without looking: Analysis pipeline for child sexual abuse datasets. *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 2189–2205).

Lee, Y. (2024). *Thorn and all tech is human forge generative AI principles with AI leaders to enact strong child safety commitments*. Thorn Blog. https://www.thorn.org/blog/generative-ai-principles/.

Lee, H.-E., Ermakova, T., Ververis, V., & Fabian, B. (2020). Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation, 34*, Article 301022.

Levine, A. S. (2023). Stable diffusion 1.5 was trained on illegal child sexual abuse material, stanford study says. *Forbes Magazine*. https://www.forbes.com/sites/alexandralevine/2023/12/20/stable-diffusion-child-sexual-abuse-material-stanford-internet-observatory/.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv [cs.CL]. arXiv*. http://arxiv.org/abs/1907.11692.

Lorenzo-Dus, N., Evans, C., & Mullineux-Morgan, R. (2023). Online child sexual grooming discourse. In *Elements in forensic linguistics*. Cambridge University Press.

Macedo, J., Costa, F., & A. dos Santos, J. (2018). A benchmark methodology for child pornography detection. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)* (pp. 455–462).

Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2019). Do GANs leave artificial fingerprints?. In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)* (pp. 506–511).

McNicholas, T. (2023). *New Jersey high school students accused of making AI-generated pornographic images of classmates*. New York: CBS. https://www.cbsnews.com/newyork/news/westfield-high-school-ai-pornographic-images-students/.

Microsoft. (2009). New technology fights child porn by tracking its "PhotoDNA. https://news.microsoft.com/2009/12/15/new-technology-fights-child-porn-by-tracking-its-photodna/.

Murphy, M. (2023). *Apps that use AI to undress women in photos soaring in use* (Vol. 3). https://news.bloomberglaw.com/artificial-intelligence/apps-that-use-ai-to-undress-women-in-photos-soaring-in-use.

Nazah, S., Huda, S., Abawajy, J. H., & Hassan, M. M. (2021). An unsupervised model for identifying and characterizing dark web forums. *IEEE Access: Practical Innovations, Open Solutions, 9*, 112871–112892.

Ngo, V., McKeever, S., & Thorpe, C. (2023). Identifying online child sexual texts in dark web through machine learning and deep learning algorithms. https://doi.org/10.21427/WFN5-RT72.

Nguyen, T. T., Wilson, C., & Dalins, J. (2023). Fine-tuning Llama 2 large Language Models for detecting online sexual predatory chats and abusive texts. *arXiv [cs.CL]. arXiv*. http://arxiv.org/abs/2308.14683.

O'Brien, M., & Hadero, H. (2023). AI-generated child sexual abuse images could flood the internet. *Now there are calls for action*. AP News. https://apnews.com/article/ai-artificial-intelligence-child-sexual-abuse-c8f17de56d41f05f55286eb6177138d2.

OpenAI. (2022). DALL·E 2 pre-training mitigations. https://openai.com/research/dall-e-2-pre-training-mitigations.

Peersman, C., Schulze, C., Rashid, A., Brennan, M., & Fischer, C. (2016). iCOP: Live forensics to reveal previously unknown criminal media on P2P networks. *Digital Investigation, 18*, 50–64.

Pendar, N. (2007). Toward Spotting the Pedophile Telling victim from predator in text chats. *Proceedings of the international conference on*. https://dl.acm.org/doi/abs/10.1109/ICSC.2007.102.

Pereira, M., Dodhia, R., Anderson, H., & Brown, R. (2020). Metadata-based detection of child sexual abuse material. *arXiv [cs.LG]. arXiv*. http://arxiv.org/abs/2010.02387.

Quayle, E. (2020). Prevention, disruption and deterrence of online child sexual exploitation and abuse. *ERA Forum*. https://doi.org/10.1007/s12027-020-00625-7

Rezaee Borj, P., Raja, K., & Bours, P. (2023). Detecting online grooming by simple contrastive chat embeddings. In *Proceedings of the 9th ACM international workshop on security and privacy analytics* (pp. 57–65).

Ringenberg, T., Seigfried-Spellar, K., & Rayz, J. (2021). Implications of using internet sting corpora to approximate underage victims. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 3645–3656). Association for Computational Linguistics.

Ringenberg, T. R., Seigfried-Spellar, K., & Rayz, J. (2024). Assessing differences in grooming stages and strategies in decoy, victim, and law enforcement conversations. *Computers in Human Behavior, 152*, Article 108071.

Ritch, F. (2022). The Artifice girl [Film]. XYZ Films https://www.imdb.com/title/tt20859464/.

Rondeau, J. (2019). Deep learning of human apparent age for the detection of sexually exploitative imagery of children [university of Rhode Island]. https://search.proquest.com/openview/053cbe215c25bc2a51095816de35bd69/1?pq-origsite=gscholar&cbl=18750&diss=y.

Sae-Bae, N., Sun, X., Sencar, H. T., & Memon, N. D. (2014). Towards automatic detection of child pornography. In *2014 IEEE international conference on image processing (ICIP)* (pp. 5332–5336).

Sanchez, L., Grajeda, C., Baggili, I., & Hall, C. (2019). A practitioner survey exploring the value of forensic tools, AI, filtering, & safer presentation for investigating child sexual abuse material (CSAM). *Digital Investigation, 29*, S124–S142.

Scanlan, J., Prichard, J., Hall, L., Watters, P., & Wortley, R. (2024). *reThink chatbot evaluation*. https://figshare.utas.edu.au/articles/report/reThink_Chatbot_Evaluation/25320859/1/files/44757073.pdf.

Seto, M. C., & Eke, A. W. (2015). Predicting recidivism among adult male child pornography offenders: Development of the Child Pornography Offender Risk Tool (CPORT). *Law and Human Behavior, 39*(4), 416–429.

Singh, S., & Nambiar, V. (2024). Role of artificial intelligence in the prevention of online child sexual abuse: A systematic review of literature. *Journal of Applied Security Research*, 1–42.

Steel, C. M. S. (2014a). *Digital child pornography: A practical guide for investigators*. Lily Shiba Press.

Steel, C. M. S. (2014b). Idiographic digital profiling: Behavioral analysis based on digital forensics. *Journal of Digital Forensics, Security and Law, 9*(1), Article 1.

Steel, C. M. S. (2021). Digital behaviours and cognitions of individuals convicted of online child pornography offences [The University of Edinburgh] https://doi.org/10.7488/ERA/1634.

Steel, C. M. S., Newman, E., O'Rourke, S., & Quayle, E. (2020). An integrative review of historical technology and countermeasure usage trends in online child sexual exploitation material offenders. *Forensic Science International: Digital Investigation, 33*, Article 300971.

Steel, C. M. S., Newman, E., O'Rourke, S., & Quayle, E. (2022a). Improving child sexual exploitation material investigations: Recommendations based on a review of recent research findings. *Police Journal*, Article 0032258X221142525.

Steel, C. M. S., Newman, E., O'Rourke, S., & Quayle, E. (2022b). Public perceptions of child pornography and child pornography consumers. *Archives of Sexual Behavior*. https://doi.org/10.1007/s10508-021-02196-1

Stelmack v. State. (2010). 58 so. 3d 874 (dist. Court of appeals. https://scholar.google.com/scholar_case?case=14871498681120218966.

Sweeney, L. (2003). That's AI?: A history and critique of the field. https://kilthub.cmu.edu/ndownloader/files/12102473.

Thorn. (2024). Safety by design for generative AI: Preventing child sexual abuse. https://doi.org/10.25740/jv206yg3793.

United States of America, V. David Tatum. (2023). Defendant, United States OF (United States v. Tatum. https://scholar.google.com/scholar_case?case=9866333615317316908.

Vitorino, P., Avila, S., Perez, M., & Rocha, A. (2018). Leveraging deep neural networks to fight child pornography in the age of social media. *Journal of Visual Communication and Image Representation, 50*, 303–313.

Vogt, M., Leser, U., & Akbik, A. (2021). Early detection of sexual predators in chats. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 4985–4999).

Vorsino, Z. (2021). Chatbots, gender, and race on web 2.0 platforms: Tay.AI as monstrous femininity and abject whiteness. *Signs: Journal of Women in Culture and Society, 47*(1), 105–127.

Whyte, C. (2020). Deepfake news: AI-enabled disinformation as a multi-level public policy challenge. *Journal of Cyber Policy, 5*(2), 199–217.

Woodie, A. (2016). Yahoo shares algorithm for identifying "NSFW" images. *Datanami*. https://www.datanami.com/2016/10/03/yahoo-shares-algorithm-identifying-nsfw-images/.

Yiallourou, E., Demetriou, R., & Lanitis, A. (2017). On the detection of images containing child-pornographic material. In *2017 24th international conference on telecommunications (ICT)* (pp. 1–5).

Yu, N., Davis, L. S., & Fritz, M. (2018). Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *IEEE international conference on computer vision* (pp. 7555–7565).

Zhang, X., Karaman, S., & Chang, S.-F. (2019). Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–6).

Internet Watch Foundation. (2024). *Pioneering chatbot reduces searches for illegal sexual images of children*. Internet Watch Foundation. Retrieved March 3, 2024, from https://www.iwf.org.uk/news-media/news/pioneering-chatbot-reduces-searches-for-illegal-sexual-images-of-children/.

Panchenko, A., Beaufort, R., & Fairon, C. (2012). Detection of Child Sexual Abuse Media on P2P Networks: Normalization and Classification of Associated Filenames. Language Resources for Public Security. In *Proceedings of the LREC Workshop on Language Resources for Public Security Applications* (pp. 27–31). https://core.ac.uk/download/pdf/38625377.pdf#page=32.

Westlake, B., Brewer, R., & Swearingen, T. (2022). Developing automated methods to detect and match face and voice biometrics in child sexual abuse videos. *Trends and Issues in Crime and Criminal Justice*, (648), 1–15. https://search.informit.org/doi/abs/10.3316/agispt.20220331064671.